Journal of the
OPTICAL SOCIETY
of AMERICA

VOLUME 32 AUGUST, 1942 NUMBER 8

Electron and Optical Microscope Interpretation of the Wall of *Pleurosigma angulatum*

D. H. HAMLY AND J. H. L. WATSON
University of Toronto, Toronto, Canada

(Received May 25, 1942)

RECENTLY numerous papers have appeared discussing the electron microscope and its applications in industry and research—Hall and Schoen,[10] Zworykin and Ramberg,[25] Mudd and Lackman,[17] Burton and Kohl,[26] Anderson and Stanley,[1] Klemperer,[13] and von Ardenne.[24] Last year the authors demonstrated how the electron and optical microscopes supplement each other in interpretation over a size range of 0.1μ to 0.5μ, where the first instrument presents a well-resolved but monoplanar image and the latter a poorly resolved but three-dimensional image (Hamly and Watson[12]). The present paper discusses the use of both instruments in interpreting the wall structure of the diatom *Pleurosigma angulatum* which can be well resolved and easily photographed in the electron microscope.

Microscopists have discussed warmly the use of diatoms and ruled gratings as test objects for many years, and have argued the practical advantages of Abbe illumination as against those of Nelson critical illumination, and the Abbe theories of image formation as against those accepted as applying in the case of non-microscopic images. Notwithstanding lively differences of opinion and failure to reach agreement, both diatoms and gratings have been used widely. Dippel,[7] Behrens,[5] Spitta,[22] Beck,[3] and others have published tabulated information on diatoms and gratings suitable for test use.

Many different diatoms have been employed, and of these *P. angulatum* has been used more commonly than any other, yet no generally accepted explanation of the wall structure has been made in the last seventy years. The various explanations which have been published may be divided into four groups. In one the beaded appearance of the wall is attributed to projecting points by Castracane (Dallinger[6]) (Fig. 1); in another to cavities as shown by Fig. 2 (Müller,[18] Flögel,[8] Smith,[21] and more recently Fritsch[9]); in another to large holes, as shown in Figs. 3 and 4 (Dippel,[7] van Heurck,[23] Beck[3]); and in the fourth to depressions with or without small holes.

All these views have been criticized as speculations, since they arose from the interpretation of diffraction images. Abbe voiced this objection (Flögel[8]), and later Dallinger[6] restated it using a purely geometrical figure as an illustration (Fig. 5). Accurate images were held to be unattainable in these very small structures since the use of apochromats of the largest available aperture (*NA* 1.6) still gave images which lacked crispness.

Such disagreements have thrown doubt on the usefulness of transparent periodic materials as test objects, and Barnard,[2] Beck,[3] and others have recommended the substitution of well-stained bacteria or sections of root tips showing the metaphase in mitosis. Some optical workers have turned to purely optical methods for testing

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

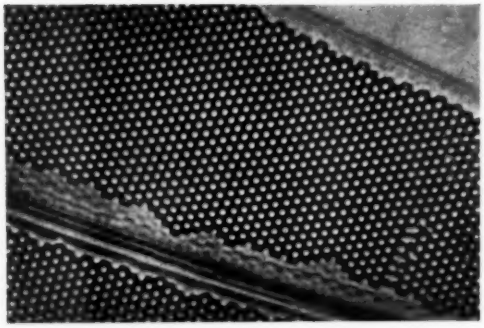


FIG. 1. *P. angulatum.* ×2300. Apochromat *NA* 1.4. Beaded structure and intercostal markings clearly shown. Resolution described as incomplete. Compare with Fig. 5. Note partial resolution outside postage stamp edge. By Dr. R. Zeiss from Dallinger (reference 6).

objectives. The Abbe silver film test slide has been used for many years, and the interferometer is a relatively new instrument capable of giving good results in the hands of opticians (Martin[16]). However, it is proposed to show that most of the published photomicrographs are concordant with a structure which has a physiological basis. Models have been made which show phenomena similar to those seen in viewing *P. angulatum* with the microscope.

TECHNIQUE

The same specimens were not compared in the electron and optical microscopes, but the equivalent result is believed to have been achieved by using in all cases *P. angulatum* specimens from the same series of Ernst Leitz test slides, in which inspection of many samples has shown very little variation in the pattern dimensions.

Optical Microscopy

Symmetrical Köhler illumination was used and all optical units were centered, oriented, and focused on a large Zeiss horizontal-vertical photomicrograph bench. A negative 3-diopter lens was inserted in the filter stop holder of the substage condenser to correct for the long bench. A ribbon filament lamp (6 v, 18 amp.) or a carbon arc lamp (55 v, 7 amp., 5 mm solid positive carbon) was used as a source of light.

Ilford microfilters 2 and 4 together pass a narrow band of blue-green light between the limits of 0.485μ and 0.555μ with a peak at 0.525μ. This band was adopted because photographs made with this illumination correspond closely in appearance with what the eye sees when the specimen is viewed with the same illumination.

An Ic (1906) Zeiss research microscope modified by the addition of a precision focusing attachment (Hamly[11]) and by arrangements for various types of substage equipment was used. While several kinds of achromatic condensers, apochromatic objectives, and compensating oculars were employed, the text illustrations were made with only two combinations. One was a modern Zeiss achromatic condenser (*NA* 1.0), a Leitz 4-mm apochromat (*NA* 0.95) with a Leitz ×10 periplan ocular, and the other was a Bausch and Lomb achromatic condenser (*NA* 1.4), with a Spencer 2-mm apochromat (*NA* 1.4), and a Leitz ×10 periplan ocular. The best images were obtained with a mechanical tube length of 162 mm for the Spencer objective and of 170 mm for the Leitz objective with the collar adjusted to the cover glass thickness.

The diatoms were mounted in realgar by the Madam[15] technique using slides and cover glasses micrometered to the limits required by the condensers and objectives. Several other mounting media were tried, and of these monobromonaphthalene (refractive index 1.66) proved most satisfactory, but the contrast obtained with this

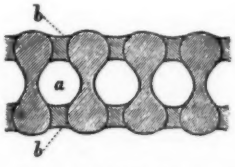


FIG. 2. *P. angulatum.* ×12000. Diagram of the chambers said to exist in the wall. The chamber *a* has an outer pore *b* and an inner pore *h*. From Müller (reference 18).

medium at the high apertures proved much less than that obtained with realgar (refractive index 2.6). Shillabeer's non-drying immersion oil was used instead of cedar immersion oil both above and below the slide.

The objectives were centered after the mirror had been correctly oriented, true alignment being indicated by the lack of image shift during the focusing of the objective and by the uniform color of the margin of the source aperture image.

Studies and serial photographs were made at several substage condenser apertures up to about *NA* 1.3. Glare was reduced by keeping the source condenser aperture image smaller than the objective field.

Oblique Abbe illumination was obtained by using a small condenser aperture (4 mm) equal to approximately *NA* 0.15, offset to the margin of the illuminating field angle and then rotated to the azimuth required by the specimen. Inspection of the back lens of the objective (apochromat *NA* 0.95) showed one non-diffracted and three diffracted beams under these conditions.

Fine focusing was carried out normally until the correct field and the mid-focal plane of the specimen were obtained and then the precision focusing attachment was clamped. Then the microscope was tapped gently, and raised and lowered by the precision control several times from out-of-focus below the specimen to out-of-focus above it to ensure that repetition of the same focal conditions was possible during photography.

After visual inspection with high power compensating oculars the Leitz ×10 periplan was inserted and the image focused on a ruled glass screen with the aid of a ×1.5 magnifying glass. The magnification was checked against an accurate micrometer slide and a steel scale, and was accurate to ±5 percent.

Eight photographs were taken on a single film using a Zeiss Multiplex camera back. Exposures were timed to smaller differences than ±10 percent. Development and fixation were carried out without cutting the film so that all changes in appearance could be ascribed to the small changes in focus which were made between exposures.

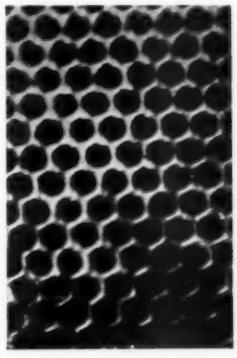


FIG. 3. *P. angulatum.* ×6300. Zeiss apochromat *NA* 1.63. Large aperture structure. Compare with Figs. 4 and 10b. Beck (reference 3) refers to this photomicrograph as showing the same structure as shown with a new high aperture dark-field condenser. By van Heurck (reference 23) from Dallinger (reference 6).

Electron Microscopy

The electron microscope used for this investigation was that designed by Prebus and Hillier,[19]

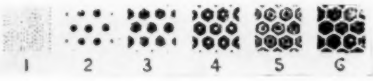


FIG. 4. *Triceratium favus.* Diagram showing changes in resolution with changes in aperture. 1, no resolution. 2–5, increasing resolution. 6, full resolution. From Beck (reference 3) who compares it to *P. angulatum.*

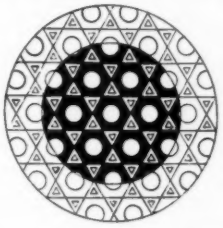
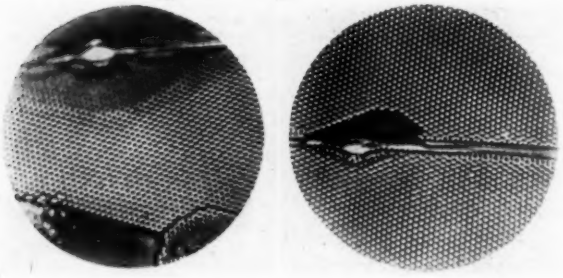


FIG. 5. *P. angulatum.* ×8200. A geometrical figure based on the Abbe diffraction experiments. From Dallinger (reference 6).

and, while mechanical changes have been effected, the performance is not essentially different from that described. This microscope is usually used for objects which are less than 0.2μ in size, but it has given satisfactory results with objects as large as 10μ. At ×2000 a suitable field for photography has a diameter of 20μ.

The electronic image has aberrations very similar to the errors found in optical images and these aberrations are given analogous names. It is imperative that these errors be as small as possible if good images are to be obtained. To ensure this certain precautions were taken. The lens system was lined up throughout, a good vacuum was maintained, the accelerating voltage of 45,000 v was kept constant within a few volts, the magnification was kept reasonably high, and the field was limited by suitable diaphragms.

The magnifications reported in this paper are those determined by optical means, and when the



a b

FIG. 6. *P. angulatum*. ×1000. Photomicrographs of the outer and inner surfaces of the wall. Compare with Fig. 10 where similar changes in appearance are obtained by changes in focus. By T. F. Smith from Dallinger (reference 6).

electron and optical photographs are compared, those of the electron microscope are adjusted to correspond to the optical magnifications. This method may produce discrepancies as high as 20 percent. This difference can arise from an error of determination, which is estimated to be at least ±5 percent in the electron microscope at ×2000, and from variation in the sizes of individual diatoms.

The electronic negatives were obtained on Ilford "Special Lantern" plates (10″×2″), with exposures of about 2 seconds and development to a gamma less than 1.2, in normal metol-hydroquinone developer (Ilford ID2). Though this gamma is not high it was necessary to use the softest grades of enlarging paper to reproduce, in the positives, the detail available in certain negatives, since the range of tones is very great (Figs. 8, 9, 13–15). Considerable care was exercised in choosing the paper because unsuitable grades give misleading impressions of dimensions and structure. The dimensions which are reported were measured on a screen from projected negatives.

Two methods were used in mounting the diatoms on the collodion covered, "electro-mesh" screen holders which are one-eighth inch in diameter and pierced by 200 holes per inch. In the first method the specimens were suspended in distilled water and a small drop of this suspension was allowed to dry on the holder. In the second method samples were mounted by simply rubbing the prepared holder over slide surface carrying the diatoms. Ordinarily the specimens lay flat on the collodion film but some were tilted. Tilt was clearly indicated by a lack of equilateral symmetry in the photographs (Figs. 17–19).

DISCUSSION

While written description has been the important evidence offered in interpretations of wall structure in *P. angulatum*, it has usually been supplemented by visual evidence of two classes, the subjective evidence of sketches and diagrams, and the objective evidence of photomicrographs.* Although there is a slight element of the subjective present, as both the original material and final print are selected for publication, even poor photomicrographs are objective evidence; each individual inspecting them may interpret them according to his own experience.

In papers which use the first class of evidence are found views favored by the biologists Fritsch[9] and Smith.[21] These interpretations are largely based on the extensive investigations of Müller[18] and Flögel.[8] Müller considered the wall of the valve as being made up of chambers opening by pores to the exterior and interior surfaces (Fig. 2). Flögel[8] considered that the outer pores were absent. Recently Beck has published sketches of the diatom *Triceratium Favus* (Fig. 4) to explain his view that the wall of *P. angulatum* has large pores, basing his interpretation on what was revealed when the valve was examined with a *NA* 1.4 objective after illumination by a *NA* 1.45–1.55 dark-field condenser. He also republishes Fig. 3 as showing the same structure.

* Good drawings of diatom valves have not been published.

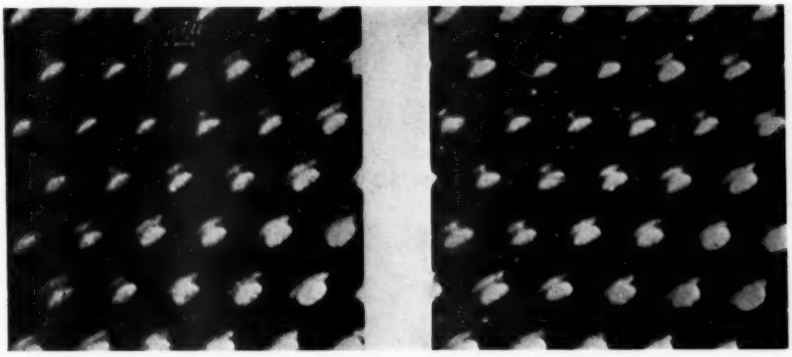
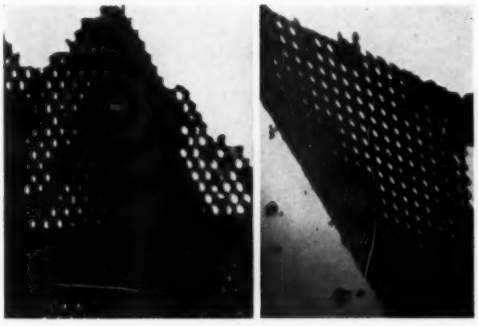


FIG. 7. *P. angulatum.* ×13500. Stereo-electronphotomicrograph of the wall looking toward the membrane with its slit-like apertures. By measuring the absolute parallax and using the authors' data×25000 swing 4° the separation of base of wall and membrane is found to be approximately 0.25μ. Basing the magnification on optical measurements this value becomes approximately 0.5μ. From von Ardenne (reference 24).

The second class of evidence occurs in the work of amateur and professional microscopists interested in structure as determinable by the microscope. These writers have rarely cited morphological processes for confirmation of their views and the interpretations are based largely on what can be seen with the microscope or in photomicrographs. The photomicrographs of *P. angulatum* can be divided into definite types, and those reproduced are probably the earliest published of each kind. Figure 1 by Dr. R. Zeiss (Dallinger[6]) shows the "beads" which are now more generally interpreted as small apertures. Figure 6 by T. F. Smith (Dallinger[6]) shows the so-called outer and inner appearance of the valve. Figure 3 by van Heurck[23] shows the large apertures referred to by Beck.[3] An outer membrane has been suspected, and appears to be definitely established by the stereoelectron photographs made by Von Ardenne (Fig. 7). These indicate the existence of a membrane pierced by regularly arranged narrow slits and holes and lying approximately 0.5μ** above the basis of the wall proper.



8 9

FIGS. 8 AND 9. *P. angulatum.* ×2500. These electron photomicrographs indicate the presence of a membrane over the wall by showing a series of slits in addition to the apertures. In Fig. 8 the membrane has been displaced by rubbing, and in Fig. 9 it is in the normal position and is seen because of the 23° tilt.

It has been argued of each of these photographs, and of others, that it is the "true" picture of the wall and that those which differ are optical artifacts. However, *P. angulatum* has been photographed optically by the techniques already described so that these photographs explain the apparently conflicting evidence of previously published photographs and form a basis for interpreting the observations of Müller,[18] Flögel,[8] and others.

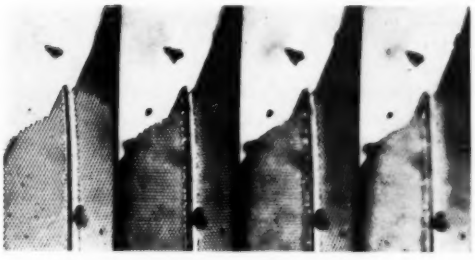
A chambered wall has not been indicated in any published photomicrographs, but the von Ardenne photographs and those shown in Figs. 8 and 9 are evidence of a thin friable membrane (van Heurck[23]) separated by a more or less continuous space from the wall proper supported by the points of the ridges of the wall. The existence of this membrane probably explains the "postage

** A value calculated from the photographic parallax, the magnification of the photograph, and the 4° stereo-rotation between photographs.

Separation of planes
= [Photographic Parallax/(Mag. of photo)×tan 4°].

b d e f

FIG. 10. *P. angulatum.* ×833. These optical photomicrographs made with axial illumination, Spencer apochromat *NA* 1.4, working aperture *NA* 1.3. B and L achromatic condenser *NA* 1.4, immersion oil above and below slide. Change in focus 0.6μ. Ilford Microfilters 2+4 with dominant wave-length of 0.53μ. Köhler axial illumination with arc. Note similarity of c to Fig. 1, and b and d to Fig. 3.



a b c d

FIG. 11. *P. angulatum.* ×833. Data as in Fig. 10, excepting working aperture *NA* 1.0 and illumination very slightly off axis. Differences in resolution caused by slightly oblique illumination rather than by change in aperture.

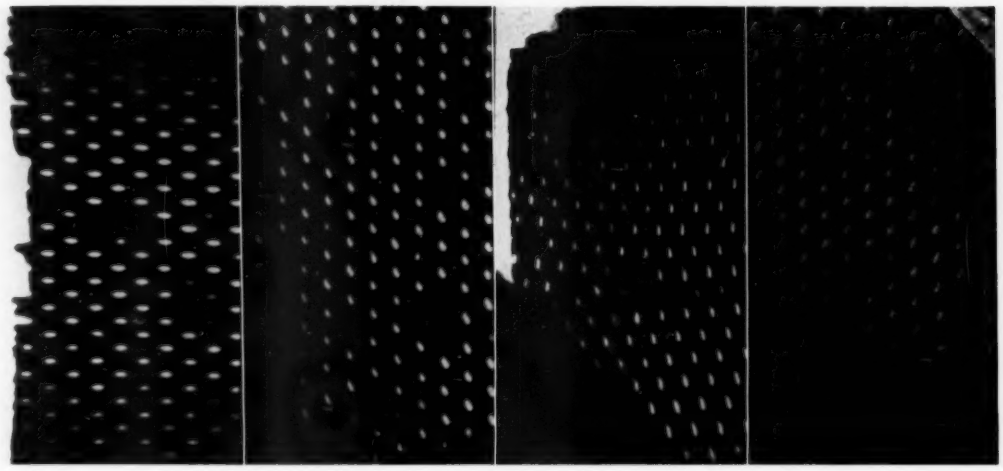
stamp edge" in Fig. 1 (not to be confused with the "broken edge" also frequently known as "postage stamp effect") and the change in appearance which takes place as a medium of one refractive index replaces that of another within this space. This change in appearance was noted by Müller[18] and Flögel[8] and used as the principal argument in support of the chamber hypothesis.

Figures 10 and 11 show a series of optical photographs which reproduce most of the details to which reference has been made. Though these photographs do not positively indicate the structure of the wall, an explanation of its character is possible when they are studied along with the oblique Abbe illumination photographs, Fig. 12 and the electron photographs, Figs. 8, 9, and 13 to 19.

The electron microscope photographs show that the wall proper has rows of elliptic apertures varying somewhat in size and in general being about 0.3μ in length with a possible range of 0.2μ to 0.4μ and a width of approximately 0.15μ. These apertures are usually spaced equilaterally by 0.7μ. In certain specimens, probably the older

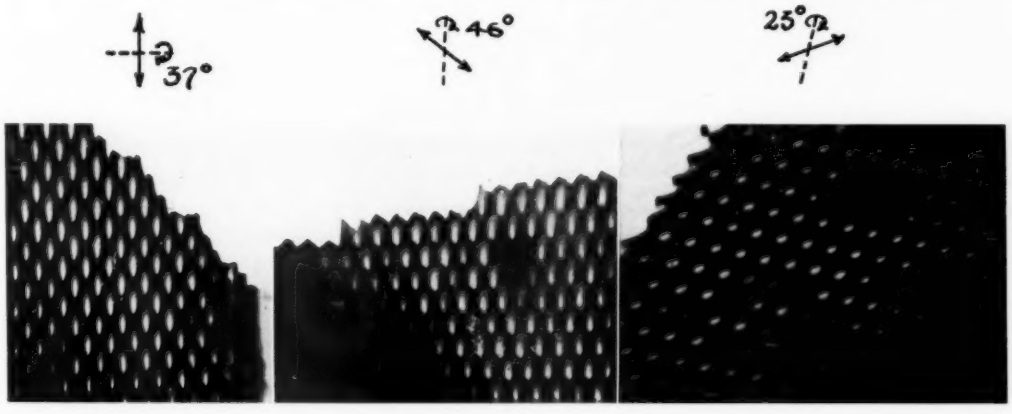
a b c d e

FIG. 12. *P. angulatum.* ×627. Oblique Abbe illumination as described in "Technique." The illumination angle to the longitudinal axis and the resolved pattern are a−60° and left oblique striae, b−90° and hexagonal block pattern, c−120° and right oblique striae, d−180° and dotted transverse striae. The same focus is shown in a, b, c, and d, while e shows a focal difference of 0.8μ.



13 14 15 16

FIGS. 13–16. *P. angulatum.* ×5000. Electron photomicrographs of specimens lying normal or nearly so to axis. Figure 13 shows diatom fractured transversely to axis. Figures 15 and 16 show diaphragm with holes in apertures.



17 18 19

FIGS. 17–19. *P. angulatum.* ×5000. Electron photomicrographs showing the change in pattern dimensions produced by tilted specimens on screen of electron microscope. The amount of tilt is indicated in degrees, the axis of tilt by - - - - - line, and the longitudinal axis by the double-headed arrow. In all cases the electron microscope was properly focused and centered. The distorted image is produced by the electrons striking the sides of the funnel-shaped apertures. Note the relation of the broken edges of valves to the longitudinal axis.

valves, these apertures are stopped by a thin diaphragm which is pierced by small holes arranged in a row parallel to the longitudinal axis of the valve. These holes are separated from each other by 0.02μ and are about 0.04μ in diameter, and four are ordinarily present unless they have been blocked by debris.

These dimensions are well within the resolving power limits of the electron microscope but far beyond the resolving power of the optical microscope. The elliptic apertures are not too small† to

† With a $\frac{5}{6}$ working aperture of a *NA* 1.4 objective using green light of wave-length of 0.56μ the limit of resolution is

$$\frac{0.61\times0.56}{\frac{5}{6}\times1.4}\mu=\frac{0.3}{1.4}\mu=0.2\mu.$$

Still smaller structures than this have been resolved in the diatoms *Amphipleura pelucida* and *Synedra delicatissima* (Spitta (reference 22), Hamly and Watson (reference 12), Köhler (reference 14).

Fig. 20

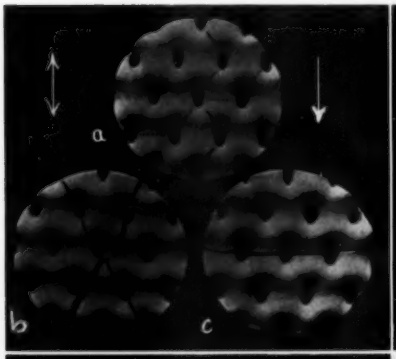


Fig. 21

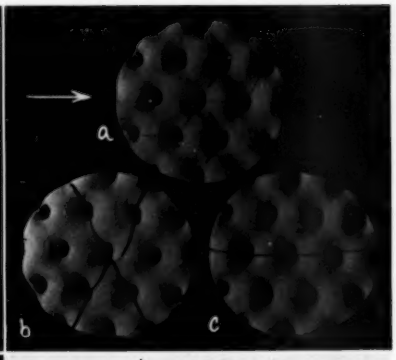


Fig. 22

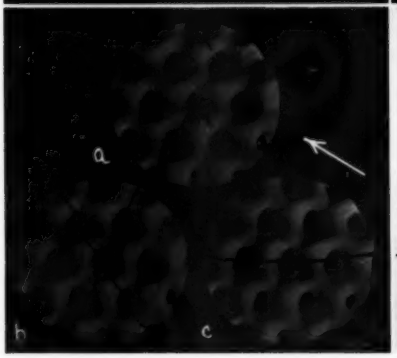
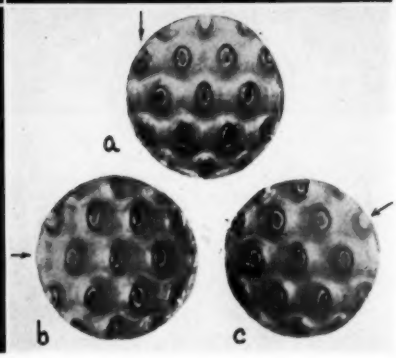


Fig. 23



FIGS. 20–22. Plaster models of *P. angulatum*. ×8800. Illuminated by incident light at 60° to the normal in the three azimuths indicated. These show: a—unbroken, with occasional lines of fracture marked; b—with right oblique fracture and line of left oblique fracture marked; c—with transverse fracture.

FIG. 23. Gelatin models of *P. angulatum*. ×8800. Supported on glass and illuminated by transmitted light at 45° to the normal in the three azimuths indicated. These are the same shape as the plaster models shown in Figs. 20–22.

be resolved by the latter, and yet there is very little evidence of their presence even when the valves are mounted in realgar. The failure of high working apertures to produce any real increase in resolution is explainable on the assumption that the openings are relatively narrow for their depth, hence increases in the refractive index of the mounting medium or changes of the working aperture can produce little improvement. Becke[4] and Rheinberg[20] have discussed the optical aspects of this problem.

Morphologists have observed that non-crystalline material in plant cell walls is laid down in regular patterns thicker in some places than in others, and it may be assumed that the siliceous wall material of diatoms takes similar forms. There is definite evidence of this in the fracturing of the wall and in peculiar effects obtained with oblique Abbe illumination.

The oblique Abbe illumination as described under "Technique" has often been used in attempting to resolve the structure of diatoms. While ineffectual in showing the true character of the aperture in *P. angulatum*, it can be used to indicate the approximate position of ridges in the walls. These are shown by illumination at right angles to the plane of the ridges when the resolving power in that plane is great enough. Figures 12 (a, b, c, d, e) show the different sets of ridges made visible by a change in azimuth of the illumination. These ridges lie in planes 30° right and left oblique, and 90° to the longitudinal axis, but not parallel to the axis. Figure 12b shows no

ridges but does give a partial pattern resolution similar to that shown in Fig. 6a and Fig. 10b.

No direct evidence of the variations in wall thickness has been obtained from the electron microscope images, but the study of valve fragments in both normal and tilted positions suggests that the thickness is probably of the order of 0.5μ, Figs. 13–19.

Fracturing occurs persistently in lines at 30° right and left oblique and perpendicular to the central axis, Figs. 8 to 19. The naturally occurring shape of fragments and the appearance seen with oblique Abbe illumination, together with all evidence obtained with the optical and electron microscopes, are believed to indicate a structure similar to that of the models shown in Figs. 20, 21, and 22.

The model, cast in plaster, fractures readily in the planes 30° right and left oblique and 90° to the longitudinal axis, but has proved difficult to break parallel to the axis and in the planes 60° right and left oblique. The diatom valve breaks occasionally in the latter planes, but the line of fracture is very rough. Figures 20, 21, and 22 show the models under three directions of illumination in order to show up the ridges. The unbroken model a shows the infrequent lines of fracture, model b shows the 30° right oblique fracture with the 30° left indicated, and c shows the transverse fracture.

When cast in gelatin and illuminated with a cone of transmitted light, and viewed on the ground glass of a camera placed directly above it, the model appears to reproduce the typical peculiarities of the diatom specimen material, Figs. 23a, b, and c. The appearance of these peculiarities is believed to indicate that the model is similar to the original specimen material. Exact reproduction of the appearance cannot be expected since the optical system for viewing the model is simpler than that of the optical microscope, with the numerical aperture smaller and the optical corrections fewer.

CONCLUSIONS

The present study using the electron and optical microscopes in conjunction with one another has indicated that it is possible to explain the apparently contradictory observations of former workers when the evidence is treated as a series of variables produced by the conditions of observation and not by actual differences in the specimen materials. The combined facts indicate that the wall of *P. angulatum* is a ridged and perforated structure covered by a thin membrane. This work has indicated, moreover, that a proper understanding of small, transparent, three-dimensioned structures in the size range of 0.1 to 1μ cannot be obtained by observation of single high power photomicrographs, but must be built up by the use of as wide a variety of visual information as possible.

REFERENCES

(1) T. F. Anderson and W. M. Stanley, "A study of purified viruses with the electron microscope," J. Biol. Chem. **139**, 325–338 (1941).
T. F. Anderson and W. M. Stanley, "A study by means of the electron microscope of the reaction between tobacco mosaic virus and its antiserum," J. Biol. Chem. **139**, 339–344 (1941).

(2) J. E. Barnard and F. V. Welch, *Practical Photomicrography* (Arnold and Company, London, 1925).

(3) Conrad Beck, "Notes on diatom structure and resolution," J. Roy. Microscop. Soc. 338–339 (1927).
Conrad Beck, *The Microscope* (R. and J. Beck, London, 1938).

(4) F. Becke, "Über die Bestimmbarkeit der Gesteingemengtheile, besonders der Plagioklase auf Grund ihres Lichtsbrechungsvermögens," Sitz. K. Akad. Wiss. (Wien) Math. Naturw. **102**, 358–376 (1893).

(5) Wilhelm Behrens, *Tabellen zum Gebrauch bei Mikroskopischen Arbeiten*, fourth edition, edited by E. Küster (Hirzel, Leipzig, 1908).

(6) W. H. Dallinger, *The Microscope and Its Revelations* (Blakiston, Philadelphia, 1891).

(7) Leopold Dippel, *Das Mikroskop und seine Andwendung. Vol. 1. Handbuch der allgemeinen Mikroskopie* (Vieweg & Sohn, Braunscheig, 1882).

(8) F. H. L. Flögel, "Über optische Erscheingung an Diatomeen," Bot. Zeits. 713, 729, 753 (1869).
F. H. L. Flögel, "Untersuchungen über die Strukture der Zellwand in der Gattung *Pleurosigma*," Arch. f. Mikr. Anatomie. **6**, 472–514 (1871).
F. H. L. Flögel, "Researches on the structure of the cell-walls of diatoms," J. Roy. Microscop. Soc. 505–522, 665–696 (1884).

(9) F. E. Fritsch, *The Structure and Reproduction of the Algae*. Vol. 1. *Bacillariophyceae*, pp. 564–651 (University Press, Cambridge, 1935).

(10) C. F. Hall and A. L. Schoen, "Applications of the electron microscope to photographic phenomena," J. Opt. Soc. Am. **31**, 281–285 (1941).

(11) D. H. Hamly, "A precision, fine adjustment for standard microscope," Science **94**, 263–264 (1941).

(12) D. H. Hamly and J. H. L. Watson, "Potentialities of optical and electron microscopes," Trans. Roy. Soc. Can. **35**, 61–65 (1941).

(13) O. Klemperer, "Electron microscopes of high magnification," *Reports on Progress in Physics*, Vol. 7 (1941), pp. 107–129.

(14) A. Köhler, Über die Feinstruktur von *Navicula* (*Pinnularia*) *nobilis* Ehbg.," Zeits. f. Bot. **22**, 442–454 (1930).

(15) H. G. Madam, "Mounting media of high refractivity," J. Roy. Microscop. Soc. 273–281 (1898).

(16) L. C. Martin, *An Introduction to Applied Optics*. Vol. I. *General and Physiological*. Vol. II. *Theory and Construction of Instruments* (Pitman & Sons, London, 1932).

(17) Stuart Mudd and D. B. Lackman, "Bacterial morphology as shown by the electron microscope," J. Bact. **41**, 415–420 (1941).

(18) O. Müller, "Über den feineren Bau der Zellwand der Bacillariaceen insbesonders des *Triceratium Favus* Ehbg. und der *Pleurosigmen*," Reichert und Du Bois Raymond's Archiv. 619–643 (1871).

O. Müller, Bemerkungen zu dem Aufsatze Dr. J. K. L. Flögel's "Researches on the structure of the cell-walls of diatoms," Ber. d. D. Bot. Ges. **2**, 487–494 (1884).

(19) A. Prebus and J. Hillier, "The construction of a magnetic electron microscope of high resolving power," Can. J. Research **17**, 49–63 (1939).

(20) Julius Rheinberg, "The black and white dot phenomena," J. Quekett Mic. Club **8**, 113–118, 301–308 (1902).

(21) Gilbert M. Smith, "The fresh-water algae of the United States," *Bacillariea*, pp. 191–270 (McGraw-Hill Book Company, New York, 1933).

(22) E. J. Spitta, *Microscopy—the Construction, Theory, and Use of the Microscope* (Murray, London, 1907).

(23) Henry van Heurck, "New objective of 1.63 *NA*," J. Roy. Microscop. Soc. 91–92 (1890) review of papers in J. Microgr. **13**, 527–528 (1889); Bull. Soc. Belg. Micr. **15**, 69–71 (1890).

Henry van Heurck, *Amphipleura pellucida* and *Pleurosigma angulatum*, J. Roy. Microscop. Soc. 103–104 (1890).

Henry van Heurck, "Structure of diatom valves," J. Roy. Microscop. Soc. 104–106 (1890).

Henry van Heurck, *Pleurosigma angulatum*, J. Roy. Microscop. Soc. 497 (1890) review of paper in Bull. Soc. Belge Micr. **6**, 10–12 (1890).

(24) Manfred von Ardenne, *Elektron-Übermikroskopie. Physik-Technik-Ergebnisse* (Springer, Berlin, 1940).

(25) V. K. Zworykin and E. Y. Ramberg, "Surface studies with the electron microscope," J. App. Phys. **12**, 692–695 (1941).

(26) E. F. Burton and W. H. Kohl, *The Electron Microscope* (Reinhold Publishing Corporation, New York, 1942).

These references are selected from a voluminous literature. Further references can be obtained from these papers or found in the same periodicals. The journals of the Royal Microscopical Society and the Quekett Microscope Club contain many papers on the structure of diatoms.

The Fluorescence of Manganese in Glasses and Crystals

S. H. LINWOOD AND W. A. WEYL
The Pennsylvania State College, State College, Pennsylvania
(Received April 20, 1942)

USE OF OPTICAL INDICATORS FOR DETERMINING THE STRUCTURE OF GLASS

WHEN ions of the transition elements are introduced into different base glasses, they exhibit optical properties which are modified by the structure of the glassy matrix. Elements of this type, therefore, can be used as indicators and they provide a powerful tool for studying the atomic structure of glass.[1] They are particularly useful, as evidence obtained by these optical indicators supplements the interpretation of x-ray diffraction patterns in a twofold way:

(1) They allow conclusions on the position and effect of the minor glass constituents, whereas x-ray analysis is naturally limited to ions which are present in relatively high concentrations.

(2) They allow one to follow structural changes of the glass caused by variation in temperature and thermal history.

Our present picture of the atomic structure of silicate glasses is chiefly based on the interpretation of x-ray diffraction patterns. It contains much of the earlier experience gained from crystal structures of different silicates (W. L. Bragg) for it has been found that the structure of crystalline and vitreous silicates is governed by the same principle. It makes use of the laws of crystal chemistry as derived by V. M. Goldschmidt and L. Pauling. B. E. Warren and his school have contributed most of the experimental evidence, and their conclusions are consistent with the earlier observations of the properties of glasses. Silicate glasses consist of a three-dimensional network of SiO_4 tetrahedra which contains the ions of the alkalies and alkaline earths in interstitial positions. With Warren we, therefore, distinguish between those cations which take an active part in forming the glass structure such as Si^{4+}, B^{3+}, and P^{5+} and those which do not exert orienting forces on the oxygen ions but merely modify the structure and contribute to the electric neutrality of the oxygen network.

The role which a cation will play when introduced into a glass depends upon different factors, the most important of which are:

(1) Ionic size.
(2) Electric charge or valency.
(3) Configuration of the outer electrons and polarization power.
(4) Temperature viscosity conditions.

Ionic size and charge determine the ionic potential, that is, the attraction force the cation exerts upon the negatively charged oxygens. High ionic potential caused by small size and strong charge favors the active role of a cation and allows it to assume the position of a network former. Large ions with low charge such as potassium or barium make the cation play the role of a network modifier.

The ionic potential alone, however, is not decisive, but the number of electrons and their distribution have to be considered. Zinc and cadmium, for instance, are much more likely to assume network forming positions than the smaller magnesium with the same charge. This tendency expresses itself in the orthosilicates of these metals which crystallize in the phenakite structure (Be_2SiO_4) containing ZnO_4 and CdO_4 groups, respectively. The magnesium orthosilicate, on the other hand, crystallizes in the structure of the olivine type where each Mg^{++} is surrounded by six oxygens.

The average coordination number decreases with increasing thermal motion of the oxygen. High temperature, therefore, allows certain ions to enter fourfold coordination even if a higher coordination would be expected to be stable at room temperature. The coordination number of a cation can be derived from its size, or to be more specific, from the relation—radius of cation : radius of anion—by simple geometrical considerations. In crystals, abrupt changes in coordination can take place by cooperative maneuvers when one modification changes into

[1] W. A. Weyl, "Investigations on the constitution of solutions and glasses by means of optical indicators," Veröff. Kaiser Wilhelm-Institute f. Silikatforschung, Berlin-Dahlem, 7, 167–203 (1935).

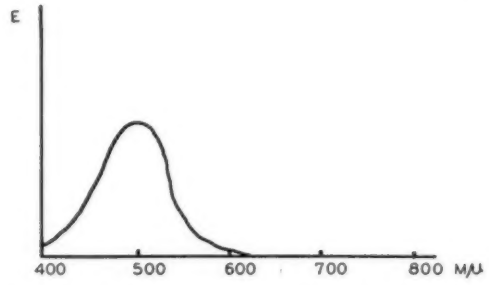


FIG. 1. Spectral absorption of cobalt in sodium-borate glass (color: pink).

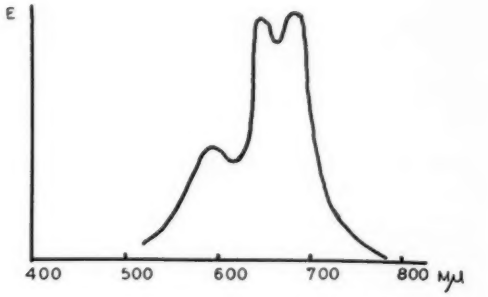


FIG. 2. Spectral absorption of cobalt in soda-lime silicate glass (color: blue).

another. In glasses, a gradual change of the average coordination number with temperature is the rule. With decreasing temperature, a high temperature equilibrium may be frozen in, for the rapidly increasing viscosity of the glass makes the atomic rearrangement more and more difficult. That explains why the temperature-viscosity relationship has to be considered an important factor for the average coordination number of a glass forming cation.

How the light absorption of a cation is affected by its position in the structure can be easily recognized from the behavior of cobalt and nickel as glass colorants.

Divalent cobalt in aqueous solutions, crystals, and glasses forms two main types of absorption centers. In a hexafold coordination, as in diluted aqueous solutions or in the periclase lattice (MgO) where it replaces an occasional magnesium atom surrounded by six oxygens, as well as in low melting borate and phosphate glasses, it gives rise to a pink color. Its absorption spectrum consists of a broad band with a maximum in the green region of the spectrum (see Fig. 1, in which extinction is plotted against wave-length). If we heat such a pink borate glass, a considerable change in the light absorption occurs and the pink color changes into a deep blue. The blue color is caused by the cobalt ion in a fourfold coordination. The higher symmetry of the CoO_4 group permits a certain fine structure of the absorption bands (Fig. 2). When agitated by thermal motion, this stable configuration oscillates chiefly as a whole unit, rather than the constituents vibrating relative to each other. This results in an absorption band originating in the near infra-red characteristic for the CoO_4 or $CoCl_4$ groups but absent in all pink cobalt compounds. In most silicate glasses, the CoO_4 group seems to be stable. The fourfold coordinated ion is also present in concentrated hydrochloric acid or when cobalt is introduced into a spinel lattice where it is known to assume fourfold coordination.

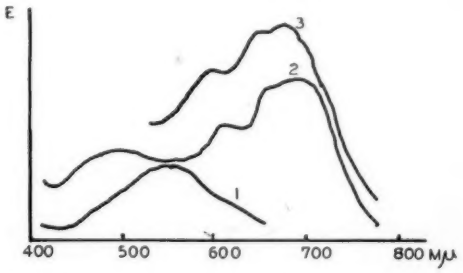


FIG. 3. Spectral absorption of cobalt in borosilicate glass as affected by temperature change (color change: pink to blue). 1 = 20°; 2 = 660°; 3 = 785°.

Using cobalt ions as indicators, we see the change in the average coordination as function of the temperature (Fig. 3). Measurements of this type give evidence of detailed structural changes, whereas the other properties such as thermal expansion or electric conductivity allow only over-all changes in the glass constitution to register.

Very similar is the behavior of nickel as a colorant in glasses. It, too, assumes fourfold coordination at high and sixfold at low temperature. Just like the cobalt, the NiO_4 group is characterized by a group of bands in the visible region of the spectrum (Fig. 4) and a band in the near infra-red. NiO_4 groups impart a purple color to glasses. The hexafold nickel resembles the hexafold cobalt, inasmuch as it causes only

one band in the visible region (Fig. 5) and has no appreciable infra-red absorption. In silicate glasses, nickel is an even better optical indicator than cobalt, for the normal soda lime glass causes the nickel to assume both positions so that a gray color results. If such a gray glass is heated, the equilibrium between NiO_4 and NiO_6 is shifted and a purple color develops, indicating that with increasing temperature, certain Ni^{++} change from hexafold into fourfold coordination. In other words, if the temperature is high enough, some of the network modifying cations can assume positions in the network proper. Most of the characteristic changes occurring in glasses in the softening range can be explained by similar atomic rearrangements.

One of the most interesting phenomena in the physical behavior of glasses is the fact that their properties are functions of the previous heat treatment. Density, refractive index, and electric properties of glasses depend upon the rate of cooling from the fused condition. This phenomenon often has been explained as a consequence of the mechanical strain induced through rapid chilling. There is ample evidence, however, that the removal of the mechanical strain is not the only process taking place when a glass is annealed, but that actually an atomic rearrangement occurs. During annealing, the atoms become much more closely packed and the refractive index of the glass increases, which is partly caused by the displacement of cations from network forming positions into network modifying positions. Optical indicators have contributed greatly to this explanation. Using nickel as an indicator in a soda lime glass, it can be demonstrated that the quenched glass has a definite purple shade which changes to a yellowish shade when the glass is properly annealed. The absorption spectra of the quenched and annealed glass allow one to estimate how many of the network forming nickel ions have been shifted into interstitial holes (Fig. 6).

It is self-evident that changes in the light absorption like those illustrated in Fig. 6 cannot be caused by a mere mechanical deformation of the outer electron positions. The importance of these considerations for the annealing of optical glass, for understanding elastic after effects and secular changes in the refractive index of glass can only be mentioned.

Not every ion of the transition group has the same sentivity to surrounding electric fields as cobalt or nickel, so that its light absorption can be used for drawing conclusions upon its environment. Divalent manganese, for instance, seems to be much less affected. All glasses which contain manganese in its divalent state are of a yellowish brown color. W. E. S. Turner and W. A. Weyl[2] found that divalent manganese

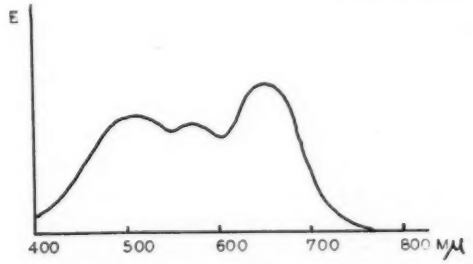


FIG. 4. Spectral absorption of nickel in potassium silicate glass (color: purple).

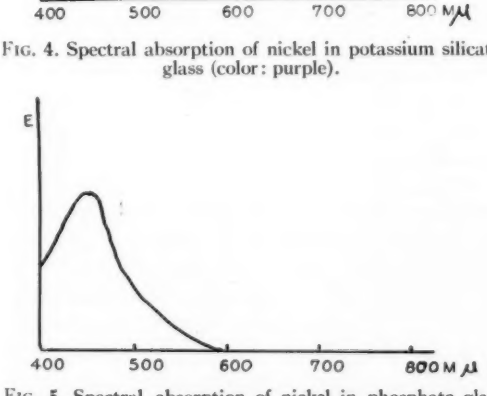


FIG. 5. Spectral absorption of nickel in phosphate glass (color: yellow).

FIG. 6. Spectral absorption of nickel in soda-lime silicate glass as affected by thermal history (color: annealed glass—yellowish gray, quenched glass—purplish gray).

[2] W. E. S. Turner and W. Weyl, "The constitution and color of glasses containing iron and manganese oxides," J. Soc. Glass Tech. **19**, 208–216 (1935).

causes a weak but rather narrow absorption band in the visible spectrum near 4250A. The narrow band is an indication that the electrons of manganese are little perturbed by the electric fields of the neighboring oxygen atoms.

This holds true not only for manganese glasses, but for crystals, too. F. A. Kröger[3] studied the light absorption of divalent manganese present in various crystal lattices and found no larger differences to exist.

Nevertheless, Mn^{++} can be used as an optical indicator, for it exhibits a characteristic fluorescence. The light emission of an ion as a rule is more affected by the surrounding electric fields than its light absorption. A study was, therefore, made on the fluorescence of manganese in different base glasses.

THE COORDINATION OF MANGANESE IN DIFFERENT GLASSES

Manganese as an activator of fluorescence in crystals and glasses has been studied since the early work of Lecoq de Boisbaudran (1887) and of Urbain. Many a crystalline material was found suitable as a matrix and the fluorescence of manganese in different media was found to vary from a bright green to yellow, orange, and to a deep red. The same is true for different glasses containing manganese and L. Bruninghaus[4] suggested as a possible explanation that the green fluorescence might be caused by the divalent manganese Mn^{++} and the red fluorescence by the trivalent manganese ion Mn^{3+}. It is true that there exists a tendency for manganese to change into the trivalent state, and nearly all glasses melted in an oxidizing atmosphere assume a purple color due to the formation of Mn^{3+}.

The equilibrium between Mn^{++} and Mn^{3+} in glass can be shifted by irradiation with ultraviolet light. This is annoying, as most technical glasses contain traces of manganese introduced as an impurity of the limestone. Sometimes a deep purple tint is developed when such a glass is exposed to sunlight.

The same process is the cause for many phototropic changes in crystals containing manganese. When the electron returns only at slightly elevated temperature, such a crystal or glass exhibits the phenomenon of thermoluminescence. When ions such as Fe^{3+} or As^{5+} are present in the glass, the electrons stand a fair chance of being caught into a more permanent position. That makes the purple tint of the glass rather stable and we call such a phenomenon solarization. Usually we attribute it to a chemical reaction of the following type:

$$2Mn^{++}+As^{5+}=2Mn^{3+}+As^{3+},$$
$$Mn^{++}+Fe^{3+}=Mn^{3+}+Fe^{++}.$$

The discoloring of manganese glass has been studied repeatedly both from the practical and scientific viewpoints.

Nevertheless, the fluorescence color of manganese and its dependence upon the glass composition have not yet been satisfactorily explained. It is possible to melt glasses free of trivalent manganese by adding reducing agents such as carbon or arsenic to the melt. Still, sometimes a green and sometimes a red fluorescence can be observed. The different fluorescence cannot be explained by the state of oxidation. Zinc borate glass has the faint reddish fluorescence characteristic of many borate glasses and M. Curie[5] found it to change into a brilliant green when devitrified. There is no reason to assume that the crystallization process should influence the state of oxidation. There is sufficient evidence that in glasses as well as in crystals, only the Mn^{++} exhibits fluorescence, so that both red and green emission bands must be attributed to the same state of oxidation.

The two types of absorption spectra of cobalt and nickel glasses have been explained on the basis of different surroundings or, to be more specific, by the different coordination of these ions. We suggest the same explanation for Mn, namely, that the change of the coordination number of Mn^{++} is responsible for the variation of the fluorescence with glass composition. Cobalt (0.82A), nickel (0.78A), and manganese (0.91A) have about the same ionic size so that it is fair to assume two fluorescent centers, one of which being Mn^{++} in fourfold, the other in

[3] F. A. Kröger, "Some optical properties of zinc silicate phosphors," Physica **6**, 764–778 (1939).

[4] L. Bruninghaus, "On the red fluorescence of glass," J. de Physique **2**, [7] 398–402 (1931).

[5] M. Curie, "Phosphorescent glasses," Trans. Faraday Soc. **35**, 114–118 (1939).

TABLE I.

GREEN	YELLOW	ORANGE	RED
$Na_2O \cdot 2SiO_2$ — $Na_2O \cdot 2SiO_2 \cdot 0.25TiO_2$			
$Na_2O \cdot 2SiO_2$ — $Na_2O \cdot 2SiO_2 \cdot 0.25Al_2O_3$			
$Na_2O \cdot 2CaO \cdot 6SiO_2$	— $Na_2O \cdot 6B_2O_3$		
$ZnO \cdot P_2O_5$	— P_2O_5		
$ZnO \cdot P_2O_5$	— $Na_2O \cdot ZnO \cdot P_2O_5$		
$ZnO \cdot P_2O_5$	— $CaO \cdot ZnO \cdot P_2O_5$		
B_2O_3	— $Na_2O \cdot 2B_2O_3$		
		$BaO \cdot P_2O_5$	
	P_2O_5 — $Na_2O \cdot P_2O_5$		
	$Na_2O \cdot P_2O_5$ — $Na_2O \cdot P_2O_5 \cdot 0.25Al_2O_3$		
	$Na_2O \cdot P_2O_5$ — $Na_2O \cdot Al_2O_3 \cdot P_2O_5$		
	$Na_2O \cdot P_2O_5$ — $2Na_2O \cdot P_2O_5$		

sixfold coordination; which of these causes a green and which red fluorescence has to be decided.

THE FLUORESCENCE OF MANGANESE AS COMPARED WITH THE COLOR OF COBALT

One way to come to a decision which fluorescence should be attributed to the fourfold and which to the sixfold coordination is to compare the change in light emission of the manganese with that of the light absorption of the cobalt. It seems only reasonable that base glasses favoring CoO_4 groups rather than CoO_6 groups would favor the same surrounding for manganese, too. The equilibrium between CoO_4 and CoO_6 cannot be expected to be identical with that between MnO_4 and MnO_6, but a certain change in a base glass should shift both equilibria in the same direction.

From the numerous glass melts studied, a few typical examples are sufficient to indicate how the fluorescence of manganese depends upon the base glass (see Table I). It was found that whenever a change in composition favors fourfold coordination of the cobalt in the corresponding manganese glass, the green fluorescence was emphasized. In low melting borates and phosphates where cobalt and nickel are likely to assume their highest coordination, the fluorescence of manganese was always orange or red.

From this information we conclude that the green fluorescence occurs whenever Mn^{++} plays the role of a network former—this means when MnO_4 tetrahedra are interlinked with the SiO_4 tetrahedra of the glass network. The minimum perturbation of the manganese ion when present in such a stable configuration allows the emission of relatively large energy quanta. The high efficiency of fluorescence is in accordance with the sharp absorption bands characteristic for the Co^{++} in the corresponding configuration.

The orange or red fluorescence, therefore, has to be attributed to those Mn^{++} ions which occupy interstitial positions where they are surrounded by a larger number of oxygens at a greater distance. The electric field lacks the symmetry of the MnO_4 group and is more subject to fluctuations due to the intense thermal vibration of the ion, in such a hole position.

INFLUENCE OF TEMPERATURE ON THE FLUORESCENCE OF MANGANESE GLASSES

Additional information for and confirmation of this explanation was obtained from the behavior of manganese glasses at different temperatures. Regardless of the composition of the glass, the following rules concerning the influence of temperature could be established:

(a) Color and intensity of green fluorescent glasses did not change to a great extent when the glass was slightly heated or when it was cooled down to the temperature of liquid air. Only above 200°C did the fluorescence begin to fade.

(b) Red fluorescing glasses show an increase in intensity when cooled, and lose their fluorescence completely when slightly heated. Some of the glasses become non-fluorescent shortly above 100°C.

(c) Glasses whose fluorescence color makes it apparent that both green and red fluorescent centers are present, show a shift towards red when cooled and towards green when heated.

We might express these results by saying that the green fluorescence has a lower temperature coefficient of intensity than the red one. This agrees very well with the role of both centers in the glass structure. The green is a part of a rather rigid network and as such, is not so much subject to the influence of thermal vibration

as the more loosely bound red fluorescing ions in interstitial positions.

CONCENTRATION QUENCHING IN MANGANESE GLASSES

In order to find the optimum concentration of manganese for our experiments, some series were investigated where the manganese concentration was varied over a wide range. There, as in other fluorescing systems such as gases, solutions, or crystals, it was observed that with increasing concentration the fluorescence first increases, then reaches a peak and decreases again. The first addition increases the number of fluorescent centers, but when a certain critical concentration is reached, quenching occurs. The true nature of this concentration quenching in crystals and glasses is not yet fully understood, but it seems that resonance phenomena are responsible which occur when two fluorescent centers are tied together by chemical or van der Waals forces. In solution, dipole association is known to facilitate the energy dissipation of excited molecules.

Glasses exhibiting the green type of fluorescence were found to be much more susceptible to concentration quenching than those with red fluorescence. In the green fluorescing silicate glasses, quenching begins when about every twentieth silicon has been replaced by manganese. With the replacement of every tenth Si^{4+} by Mn^{++} the fluorescence at room temperature has practically disappeared.

In the orange and red fluorescing glasses, the limiting concentration was not due to quenching but to devitrification. Up to 15 percent MnO was introduced into alkali borates and still they showed distinct orange fluorescence.

A glance at the schematic picture of the atomic arrangement in a sodium silicate glass (Fig. 7) as determined by B. E. Warren[6] makes us see the reasons why the two types of Mn^{++} centers respond so differently to variations in concentration.

In replacing an increasing number of Si^{4+} by Mn^{++}, soon a concentration is reached where two Mn^{++} become direct neighbors. That leads to the formation of Mn−O−Mn bonds, which are the strongest bonds possible between two Mn^{++} ions in the glass. That this configuration cannot fluoresce is indicated by the complete lack of fluorescence of pure manganese oxide.

At the other hand, even if all the network modifying sodium ions would have been replaced by Mn^{++}, no such intimate contact between two Mn^{++} will be established. All the oxygens are bound to the network formers primarily, which makes it impossible that two network modifying cations are tied together by one oxygen.

The suggested structures of the green and red fluorescing Mn^{++} centers seem to agree also with their individual susceptibility to concentration quenching.

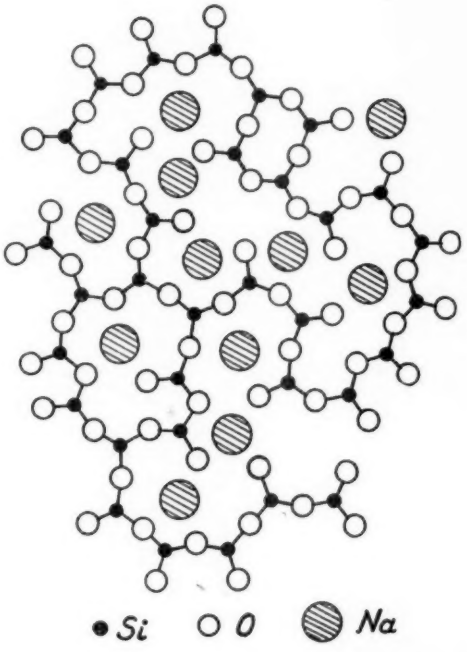


FIG. 7. Atomic arrangement of soda silicate glass, according to Warren and Biscoe. To correspond to two dimensions, the silicons are shown bonded to three oxygens rather than to four. Each silicon atom is tetrahedrally bonded to four oxygen atoms, the fourth being directly above or below the silicon atom and therefore not shown.

[6] B. E. Warren and J. Biscoe, "Fourier analysis of x-ray patterns of soda-silica glass," J. Am. Cer. Soc. **21**, 259–265 (1938).

THE FLUORESCENCE OF MANGANESE IN CRYSTALS AND THE CRYSTAL CHEMISTRY OF SILICATE PHOSPHORS CONTAINING DIVALENT MANGANESE

From our investigation on the fluorescence of divalent manganese in glasses, we deduced that the green fluorescence is caused by fourfold coordination leading to MnO_4 groups. The comparatively weak thermal vibration within such a group allows large quanta to be emitted and makes the fluorescent center rather inert towards changes in temperature. Red fluorescence, on the other hand, was found to be caused by Mn^{++} which take their places in interstices where they are more subject to perturbation. This is partly due to the more spacious and less symmetrical arrangement of the surrounding oxygens, partly to the Mn—O bonds being weaker and allowing stronger interatomic vibrations.

It is interesting to compare this proposed explanation with the well-established facts concerning the fluorescence of manganese in crystalline phosphors, especially those of the zinc orthosilicate type. In the past, crystalline phosphors have been studied much more thoroughly than glasses, for they are used commercially in fluorescent lamps and television tubes. In silicates, manganese represents the chief activator and just as with glasses, all types of emission spectra can be produced when the host lattice is changed. There is still an open question whether the Mn^{++} responsible for the fluorescence of the crystal has to assume a place in the lattice, or if it occurs in interstitial positions. Crystal chemical considerations, however, allow one to draw reasonable conclusions on what the manganese will do under various conditions.

THE GREEN ZINC ORTHOSILICATE PHOSPHOR

Green fluorescence is obtained when manganese is introduced into willemite, Zn_2SiO_4, where it replaces an occasional zinc ion. As the zinc ions in the structure of this mineral are surrounded by four oxygens, we can expect that the manganese, too, has the same environment. Indeed F. A. Kröger[3] found a homogeneous phase to form when mixtures of zinc orthosilicate with manganese orthosilicate were heated to 1200°C. Up to 50 percent of the zinc ions could be replaced by manganese ions without impairing the structure. The green fluorescence of willemite can, therefore, be ascribed to the Mn^{++} in fourfold coordination which makes it analogous to the center causing the green fluorescence in glasses.

THE RED AND YELLOW PHOSPHORS

A. Schleede and A. Gruhl[7] found that whenever a zinc orthosilicate activated by manganese was fused and quenched very rapidly, an amorphous or vitreous product was obtained which showed a weak red fluorescence, whereas less chilling leads to a product with yellow fluorescence. The structure of the yellow product was found to be crystalline, different from the willemite, but it could not be identified by these authors. G. R. Fonda,[8] who repeated these experiments, confirmed their findings and made it probable that the yellow phosphor is connected with a cristobalite structure.

There can be no doubt that the red phosphor obtained by extreme chilling is a true glass consisting of a mixed ZnO_4-SiO_4 network which contains some of the zinc and all the manganese ions in network modifying positions. That explains why the fluorescence is red rather than green.

The yellow phosphor is crystalline and cannot be explained by a vitreous phase. It might well be that when quenching zinc orthosilicate from a fused condition, Schleede obtained a product which contained both the yellow crystalline phosphor and the red glassy phosphor. The formation of the vitreous phase is, however, excluded when the yellow phosphor is synthesized by solid phase reaction.

In order to learn more about the kinetics of phosphor formation, E. Tiede and E. Weiss[9] devised a method of slow synthesis where the constituents of a phosphor were allowed to diffuse and react at relatively low temperature

[7] A. Schleede and A. Gruhl, "X-ray observations on fluorescent zinc silicate," Zeits. f. Elektrochemie **29**, 411–412 (1923).

[8] G. R. Fonda, "The yellow and red silicate phosphors," J. Phys. Chem. **43**, 561–577 (1939).

[9] E. Tiede and E. Weiss, "On the structure of phosphorus and its relation to atomic questions," Ber. d. D. Chem. Ges. **65**, 364–372 (1932).

between 200° and 800°C. D. Dobischek,[10] who applied these methods to the synthesis of willemite activated by manganese, found that the yellow phosphor is formed as the first reaction product, and that the normal green phosphor required longer times and higher temperature. Exposing the mixture of zinc oxide, silica gel, and a manganese compound to temperatures above 250°C, an orange fluorescence was noticeable after several hours. Up to 700°C, only the yellow fluorescence could be obtained, and it required 750°C or higher to synthesize the green phosphor by solid phase reaction. The optimum temperature for the green type was found to be about 900°C. In this temperature region, the formation of a vitreous phase excludes itself, as the lowest fusing eutectic mixture between zinc oxide and silica has been found to melt at 1432°C.

The phase diagram as determined by E. N. Bunting does not give evidence of a second modification of zinc orthosilicate, so the willemite structure must be considered to be stable from its melting point down to room temperature.

As, however, A. Schleede and A. Gruhl, as well as G. R. Fonda, found another crystal structure to exist, the following explanation is suggested:

The spacious structure of cristobalite, that is, the high temperature form of SiO_2, is a favorite structure of many compounds at high temperature. It tolerates substitutions which are not found in less spacious structures. T. Barth and E. Posnjak[11] made a special study on silicate structures of the cristobalite type and found that half of the silicon ions can be replaced by aluminum ions. In order to preserve electroneutrality, for each Al^{3+} substituting for a Si^{4+} one Na^{+} has to be placed in the holes of the lattice. If half of the Si^{4+} are replaced by $Na^{+}Al^{3+}$, the alpha-carnegieite results. Barth and Posnjak went even further and said that Si^{4+} can also be replaced by Ca^{++} only that now two additional sodiums have to be brought into the open structure for each calcium ion introduced. Such a substitution leads to the compounds Na_2CaSiO_4 which, like alpha-carnegieite, possesses cristobalite structure. The length of the edges of the unit cells are:

cristobalite, SiO_2	7.17A
alpha-carnegieite $NaAlSiO_4$	7.37A
sodium-calcium orthosilicate Na_2CaSiO_4	7.48A.

The entrance of the relatively large sodium ion causes only a minor change in the dimensions of the unit cell which in itself is a strong indication for the spacious character of this structure.

$$\begin{array}{ccc} & \swarrow Si_2^{4+}O_4^{--} \searrow & \\ Na^{+}Al^{3+}Si^{4+}O_4^{--} & & Na_2^{+}Ca^{++}Si^{4+}O_4^{--}. \end{array}$$

Zinc represents a much more suitable substitute for silicon than the larger calcium ion, even if Ca^{++} is known to occur as an active center with fourfold coordination in certain high temperature compounds such as tricalcium silicate. It, therefore, seems only logical that in the cristobalite structure a part of the Si^{4+} can be replaced by Zn^{++} when only for each Zn^{++} ion entering the structure an additional Zn^{++} or Mn^{++} is placed in the interstitial holes. The resulting zinc silicate of cristobalite structure, then, has two different types of zinc ions, one in fourfold coordination taking the place of a silicon and another in higher coordination filling some of the holes of the spacious structure. As Mn^{++} (0.91A) is slightly larger than the Zn^{++} (0.83A), it will replace the zinc in the hole positions preferably. From our previous experience, we should expect such a phosphor to fluoresce with a lower frequency than one having the willemite structure. That the emitted light is yellow or orange and not red as in the glass explains itself by the more symmetric surrounding and the rigidity of the crystal structure as compared with the more flexible structure of the glass.

Formation of this hypothetical modification of zinc silicate can be expected—according to Ostwald's law—whenever the melt is cooled too rapidly to allow the stable green form to crystallize, but not rapidly enough to cause glass formation.

It might also form as an intermediate unstable product when zinc oxide and silica are allowed to react at low temperature. This is especially true when silica is used in the form of silica gel, for this amorphous product changes into cristo-

[10] D. Dobischek, "Properties of luminescent glasses," Thesis, University of Berlin (1934).

[11] T. Barth and E. Posnjak, "Silicate structures of the cristobalite type," Zeits. f. Krist. **81**, 376–385 (1932).

balite when heated. The cristobalite could act as a nucleus and favor the crystallization of the unstable modification of zinc silicate. According to T. Barth and E. Posnjak, cristobalite can form mixed crystals with alpha-carnegieite and Na_2CaSiO_4 so that mixed crystal formation would also be expected between cristobalite and the cristobalite-like modification of Zn_2SiO_4. Indeed, G. R. Fonda[8] found that excess of silica over the orthosilicate ratio is favorable to the formation of the yellow phosphor. This excess of silica at the temperature in question forms cristobalite, thus providing nuclei and simultaneously stabilizing the hypothetical zinc silicate by mixed crystal formation.

THE INFLUENCE OF IMPERFECTIONS IN THE CRYSTAL STRUCTURE ON THE FLUORESCENCE OF MANGANESE

The assumption of an unstable form of zinc silicate having cristobalite structure and forming from the fused condition as well as by solid phase reaction explains the green and yellow fluorescence of zinc silicate phosphors, without abandoning the rigid condition that the fourfold coordination is responsible for the green and a higher coordination such as six for the yellow and red fluorescence.

There is no doubt that this explanation also suffices for other oxygen lattices containing manganese as an activator. E. Tiede and R. Piwonka,[12] for instance, found red fluorescence when aluminum oxide was activated with manganese and J. T. Randall[13] found the same to be true for magnesium silicate. Al^{3+} and Mg^{++} both have sixfold coordination so that the same can be expected for the Mn^{++} as an occasional substitute in these phosphors.

Nevertheless, utmost precautions should be exercised in applying such a rigid rule to crystals. The real reason why certain configurations give rise to the emission of lower frequencies or lower intensities is not so much the higher number of oxygens surrounding the excited ion, as the lower symmetry of the electric field and the increased probability of energy dissipation by collisions. The influence of temperature on fluorescence gives evidence that the fluorescence center can stand only a certain maximum of thermal agitation. In order to make fluorescence possible, collision with the neighboring atoms has to be avoided during the lifetime of an excited ion. The probability that such a collision takes place depends upon the temperature of the lattice and the atomic surrounding of the excited ion. In a crystal lattice, the uniformity of temperature or energy distribution is seriously disturbed by flaws. There are always certain places in a lattice which through missing constituents or displaced ions possess higher energy than the average. These flaws are responsible for many properties of a real crystal as compared with those of the ideal structure. Flaws often represent the active centers for catalysis and many catalyzers become inactive when the structure approaches that of the ideal crystal.

The role of flaws or "Lockerstellen" has been discussed extensively by A. Smekal and his school. Smekal[14] also used this picture for explaining phosphorescence and photoconductivity and, recently, G. R. Fonda[15] based his picture of the decay behavior of phosphors on the same explanation. G. R. Fonda assumes that a certain amount of manganese ions have places close to imperfections in the crystal and, therefore, exhibit properties different from those in the more symmetrical lattice positions. The influence of flaws or imperfections in the crystal on the light absorption is well-known from the manufacture of pigments. In order to obtain brilliant colors having relatively sharp absorption edges, the solid phase reaction has to be carried out at high temperature, thus facilitating atomic rearrangement and favoring the formation of flawless crystals. In most cases, fluxes are used which not only accelerate reaction of the batch constituents but also give the reaction product a better chance to crystallize and free it from imperfections. The same rules which govern the synthesis of pigments and stains can be

[12] E. Tiede and R. Piwonka, "Alumina as a basis for fluorescent systems," Ber. d. D. Chem. Ges. **64**, 2252–2258 (1931).

[13] J. T. Randall, "The fluorescence of compounds containing manganese," Proc. Roy. Soc. **A170**, 272–293 (1939).

[14] A. Smekal, "On a molecular theory of strength," Physik. Zeits. **27**, 837–844 (1926); **39**, 228–230 (1938).

[15] G. R. Fonda, "Phosphorescence of silicate phosphors," J. App. Phys. **10**, 408–420 (1939).

applied to crystalline phosphors, and brilliancy of the color and efficiency of the phosphor deserve to be treated alike.

The use of fluxes for the synthesis of crystalline phosphors of the sulphide or silicate type and their function in recrystallizing the phosphor is well understood. Depending upon the nature and the use of the flux, there is a chance of its being entrapped and causing imperfections. Indeed, the most efficient phosphors are obtained without the use of fluxes at relatively high temperature and sufficiently long heat treatment. The number of flaws in equilibrium with the crystal lattice increases with increasing temperature. As both influences, rate of removal of flaws, and number of flaws in equilibrium counteract each other, an optimum temperature results for the preparation of each phosphor.

This influence of flaws has to be considered, whenever an explanation is to be given for the change of light emission as the synthesis of a phosphor proceeds. D. Dobischek[10] found that a preformed zinc silicate gives rise to yellow fluorescence when it is activated by heating it with manganese salts below 700°C. Such activation through a diffusion process was possible only when the zinc silicate was produced at relatively low temperature, which means when its crystal structure was far from being perfect. Zinc silicate obtained between 1100° and 1300°C could not be activated any more at 700°C for it did not contain enough flaws to allow the manganese to diffuse into its lattice within reasonable times. When activated at higher temperature only the green fluorescence could be obtained. The adsorption of Mn^{++} at the surface of the crystal, or its migration into flaws, represents a reaction of low activation energy which can take place below 700°C with reasonable speed. The formation of the green phosphor, on the other hand, requires a much higher activation energy, for it is based on a place interchange of manganese and zinc ions and, therefore, involves breakage and formation of the strongest bonds in the lattice.

In this respect there is a fundamental difference between the formation of zinc silicate and a zinc sulphide phosphor. The activation of zinc sulphide can be accomplished at much lower temperature, for the activating ion is not forced to interchange places with zinc ions. The structure of the zinc sulphide consists of sulphur tetrahedra, but only half of them have zinc atoms in their centers, the other half being empty. In the sulphide phosphor even a perfect crystal structure provides an easy way for the activator to diffuse into the lattice, and the only limitation is the size of the activating ion. At 300°C copper migrates into the zinc sulphide lattice within a few hours, causing the well-known green ZnS-Cu phosphor. Its light emission can be modified towards yellow by substituting Cd for part of the Zn. According to E. Tiede and E. Weiss[9] the reaction:

(green ZnS-Cu phosphor) + CdS
= (yellow ZnS-CdS − Cu phosphor)

does not take place below 650°C. This reaction resembles more the formation of the green willemite phosphor, as it involves the strongest bonds in the lattice.

Considering the important role of flaws in silicate phosphors, we should make the explanation of the yellow or red fluorescence broader and not emphasize the number of surrounding oxygens only. It probably corresponds better with the facts when we say that the long wave emission of the Mn^{++} occurs whenever the ion is brought into positions of higher kinetic energy, either interstitial holes or flaws in the crystal structure.

Whether or not the manganese ion itself represents a flaw depends obviously on the size of the ion which it replaces. If part of the zinc is replaced by the larger cadmium (1.03A), a serious distortion of the lattice must be the consequence. Even when the over-all symmetry of the lattice is not impaired, such mixed crystal formation increases the thermal vibration and shifts the fluorescence toward longer wavelengths. This does not necessarily indicate a complete change of the manganese position, but that the symmetry around each Mn^{++} is disturbed by such substitutions.

Replacing zinc by cadmium leads to cadmium orthosilicate, which, since the Cd^{++} is larger than Zn^{++}, has a wider crystal lattice. A manganese ion, therefore, fits in less tightly in Cd_2SiO_4 than in Zn_2SiO_4, so that its position resembles

somewhat that in an interstitial hole where it is subject to stronger vibrations.

In this respect, an observation of H. W. Leverenz and F. Seitz[16] is extremely interesting. These authors found that the replacement of Si^{4+} (0.39A) by the somewhat larger Ge^{4+} (0.44A) has a similar effect to the replacement of part of the zinc by cadmium. The fluorescence of the Mn^{++} in zinc orthogermaniate, Zn_2GeO_4, too, is more yellowish green. Both replacements Zn^{++} by Cd^{++} and Si^{4+} by Ge^{4+} have the same effect upon the crystal lattice; they widen the structure and give the Mn^{++} a greater chance to vibrate.

Replacing Zn^{++} by the much smaller Be^{++} causes a still greater distortion of the willemite lattice. It is true that phenakite Be_2SiO_4, willemite Zn_2SiO_4, and troostite $(Zn, Mn)_2SiO_4$ are isomorphous and form mixed crystals, but the infra-red spectra of these minerals, as determined by C. L. Schaefer, F. Matossi, and K. Wirtz[17] show considerable differences, indicating that Zn^{++} and Be^{++} are far from being equivalent. The beryllium-zinc silicate phosphors are anomalous mixed crystals. The distribution of Be^{++} and Zn^{+} might be statistical or follow a certain pattern, resembling the superstructure of alloys. A new symmetry superimposed on the symmetry of the willemite structure necessarily affects the fluorescence color of the manganese. According to H. W. Leverenz and F. Seitz,[16] the phosphors of the zinc beryllium silicate type are particularly sensitive to thermal treatment, which might be explained by the fact that their thermal history decides whether a statistical distribution or a superstructure is favored.

The similarity in the behavior of manganese as an activator in glasses and crystals is also expressed in the influence of temperature.

G. R. Fonda[18] determined the fluorescence of a group of zinc silicate phosphors over a wide range of temperature. The green type was only little affected when the temperature was raised from that of liquid air to about 400°K. Phosphors which emitted orange and red light showed a steady decline in intensity when the temperature was raised.

Despite this similarity, there is still a fundamental difference between the optical properties of glasses and crystals as affected by temperature changes. Light absorption as well as fluorescence of crystals and glasses at room temperature is influenced by thermal vibration. The rapidly changing electric fields cause a broadening of the spectrum bands. The absorption and fluorescence of uranyl groups in glasses and crystals is, therefore, always more or less smeared out and differs from the sharp spectrum of the uranyl salts at the temperature of liquid air. At this low temperature the uranyl group in a crystal is surrounded symmetrically. In a glass under the same condition, no symmetry exists and the spectrum is the result of a superposition of many different spectra originating from slightly different absorption or fluorescence centers. Becquerel found the broadening of an absorption band of crystals with increasing temperature to follow a simple square root law. In glasses, however, such a simple relationship cannot be expected.

[16] H. W. Leverenz and F. Seitz, "Luminescent materials," J. App. Phys. **10**, 479–493 (1939).

[17] C. L. Schaefer, F. Matossi, and K. Wirtz, "The infra-red reflection spectra of silicates," Zeits. f. Physik **89**, 210–233 (1934).

[18] G. R. Fonda, "Characteristics of silicate phosphors," J. Phys. Chem. **43**, 561–577 (1939).

Direct Spectrographic Determination of Copper and Manganese in Dyestuffs*

G. L. Buc, G. L. Royer, and A. R. Norton
Calco Chemical Division, American Cyanamid Company, Bound Brook, New Jersey
(Received May 6, 1942)

THE determination of traces of metallic constituents in organic materials by spectrographic methods is not new. However, in most of the methods which have been reported in the literature, the organic material is first destroyed either by an ashing process or by a wet oxidation procedure. Spectrographic methods are rapid; if they can be used directly, they save a great deal of time. The ashing or wet oxidation processes take considerable time for the destruction of organic materials. For this reason it was felt that spectrographic methods would have to be applied directly to the organic materials if they were to compete with the present chemical methods. This paper discusses the method and the results that we have obtained in applying spectrographic procedures directly to organic materials such as sulfur dyestuffs for the determination of the metallic constituents copper and manganese. Traces of these elements have particular significance in the dyes that are applied to fabrics which are to be rubber coated.

APPARATUS

In our spectrographic procedures we have used a Hilger medium quartz spectrograph and Hilger arc support. Since we have not made any marked changes in this apparatus, it is not necessary to describe this in any greater detail. The densitometer which we are using for reading the photographic plates is that made by the Applied Research Laboratories and again is the standard apparatus sold by them. The step sector which we are using was built in our own shop, but its construction is very similar to that which can be obtained on the market. In use it is placed as close as possible to the slit of the spectrograph. A d.c. generator capable of furnishing 10 amp. at 250 volts is the source of power, and rheostats make it possible to decrease the amperage so that in actual operation the current may be varied between 5–10 amperes. In use, the positive side of the generator is connected to the lower electrode.

REAGENTS AND MATERIALS

Dow purified 6.4 mm ($\frac{1}{4}''$) spectrographic carbons (Dow Chemical Company, Midland, Michigan) are used to prepare the lower electrodes. They are cut into 25-mm (1″) lengths and drilled to the depth of 5.5 mm ($\frac{7}{32}''$) by use of a 3.2-mm ($\frac{1}{8}''$) drill. These short 25-mm (1″) lengths are held in a special water-cooled holder, which is equipped so that they may be inserted and removed rapidly. The upper electrodes are prepared from 3.2-mm ($\frac{1}{8}''$) National spectrographic carbons (National Carbon Company, Cleveland, Ohio). These rods are broken into lengths of 65 mm ($2\frac{1}{2}''$). To remove the last traces of copper from these electrodes, they are preburned in pairs in the regular arc holder by running for a period of two minutes at an arc current of 9 amperes with a separation distance of about 10 mm. The water-cooled holder makes it possible to use these same electrodes immediately for analysis or they may be removed and inserted later if desired.

The spectrographic plates used are Eastman spectrographic plates, Type II-O, 4×10″ size. Development is carried out in Eastman D-19 developer and the plates are washed, fixed, and finally dried by an infra-red drying lamp in a current of air.

The internal standard powder is prepared from Baker's C.P. sodium chloride which has been tested and is known to contain less than 5 parts of copper per million, Baker's thallium nitrate, and Baker's silver nitrate. The mixture is prepared by weighing 3 grams of sodium chloride and 0.225 gram of thallium nitrate. To this is added 3.23 ml of a 280-p.p.m. silver solution, and the entire mixture is evaporated to dryness and ground to insure an intimate mixture of all the components.

* Presented at the Tenth Summer Conference on Spectroscopy and Its Applications at Massachusetts Institute of Technology, Cambridge, Massachusetts, July 20–22, 1942.

PROCEDURE

One (1.000) g of the sample and 0.100 g of internal standard powder were ground together in a mortar for two minutes. Approximately ten milligrams (8–12 mg) of the mixture was then placed in the crater of a preburned electrode. A small dipper was used to measure the quantity and to transfer the mixture to the electrode.

The material was then arced at a current of 6.2 amperes in the direct current arc until the sample was completely consumed. The inter-electrode distance was maintained at 10 mm by means of an image of the arc cast on a wall by a lens. The disappearance of the typical yellow flame of sodium, a distinct change in the sound of the arc, and a rapid falling off of the arc current are simultaneous indicators of complete consumption of the sample.

The sample exposure was made through four steps of the step sector on an Eastman II-O spectrographic plate. Six sample exposures and a calibration exposure through all eight steps of the step sector were usually placed on one plate. The plates were processed as recommended by the manufacturer.

The concentrations of copper and manganese were obtained from analytical curves following the usual internal standard method. Silver was used as the internal standard for copper and thallium was used for manganese. The lines used were Ag 3280.68, Cu 3273.96, Tl 2580.14, and Mn 2576.10.

In selecting the internal standard elements essentially three criteria were used.

(1) The internal standard element must be burned out of the arc when the sodium flame disappears. Both copper and manganese were consumed before the sodium flame disappeared. This was tested by moving plate studies.

(2) The internal standard element must have a line near the element for which it is to be used. This makes for convenience when reading the plate and also guarantees that the characteristics of the plate are the same for both the unknown and the internal standard element lines.

(3) The internal standard element must not be one normally encountered in the material to be analyzed. These criteria appear to be sufficient if one is satisfied with a probable error in the neighborhood of fifteen percent.

EXPERIMENTAL DATA AND SETTING UP OF ANALYTICAL CURVE

In setting up the analytical curves it was necessary to start with material of low copper and manganese content. The material chosen was a dye in which no manganese and very little copper could be detected spectrographically. To this material copper and manganese in known amounts were added. The material was then run in the manner previously described. Each concentration was run ten times. The average ratio of intensity of the unknown to the internal standard line was used to plot the analytical curve.

Extrapolating this curve as a straight line gave a value of 3 p.p.m. for the copper content of the base dye. The analysis of this dye by the chemical method was consistent with this value. In the case of the chemical method no manganese was detected. The chemical method used for copper was an estimation of the intensity of color produced in a solution of sodium diethyl dithiocarbamate. Manganese was determined by oxidation to permanganate by potassium periodate and estimated colorimetrically by comparison with a standard similarly prepared. In both chemical methods the organic matter was first destroyed by ashing in a muffle furnace at a controlled temperature to avoid loss of copper.

The range of concentration chosen was from twenty to one-hundred parts per million in the case of copper and from five to twenty-five parts per million in the case of manganese. Within these ranges of concentrations the probable error for either copper or manganese was ± 13 percent. This was based on sixty measurements involving the two elements run at three concentrations, ten measurements being made for each element at each concentration. A dyestuff sample, not the dye used to set up the analytical curves, was run in sextuplicate and the probable error in this case was ± 7 percent. The average of the six determinations gave 49 parts per million as compared to 60 parts per million for a single chemical determination.

The actual working time per sample for a spectrographic analysis for both copper and manganese based on a series of fifty samples was one half-hour. The same actual working time per sample can be maintained in a series of six

samples. The actual working time by chemical methods was one hour per sample, provided a number of samples were analyzed at one time. If a single sample is to be run the elapsed time by the spectrographic method is two and a quarter hours; by the chemical method the elapsed time is about eight hours for a single sample due to a long ashing period. No ultra-fast plate processing methods were used; with their introduction a slight decrease in the working time per sample and a large decrease in the elapsed time for a single sample could be attained. It should be further noted that in this spectrographic method both copper and manganese are determined on the same sample at the same time. In the chemical method these two elements have to be determined separately by different procedures on rather large samples which have to be decomposed by long ashing procedures.

American Emergency Standard Photographic Exposure Computer

A STANDARD Photographic Exposure Computer has recently been prepared by the American Standards Association at the request of the United States Navy. An edition of this Standard has been prepared for distribution to the general public. This edition is printed from the same plates as those prepared for the Army and Navy, but differs in that the pages have been spirit-varnished instead of being laminated with acetate sheet, and the cover material is somewhat lighter in weight than in the military edition. The civilian edition will be available to the general public at a cost of $1 per copy. Further information regarding the Standard can be obtained from Mr. John W. McNair, Secretary, Z38, American Standards Association, 29 West Thirty-Ninth Street, New York City.

A Long Optical Path in a Restricted Space

H. R. KRATZ AND J. E. MACK
University of Wisconsin, Madison, Wisconsin and Princeton University, Princeton, New Jersey
(Received June 6, 1942)

An instrument is described for attaining a long optical path by repeated traversal between a concave spherical mirror and a totally reflecting prism. Its properties are outlined and compared with those of other devices for similar purposes. The number of traversals is in principle unlimited (we have used ours in practice for several numbers of traversals up to twelve) but the angular spread of the final beam is less than that of the first traversal by an amount that might become serious for a very large number of traversals. The final image, which is almost free of aberration, is as large, and in favorable cases almost as bright, as the source; the prism yields greater brightness than an all-mirror device at short wave-lengths, but less at long wave-lengths. The optical adjustments are simple, even for many traversals. The optics of the instrument is considered in detail, with special attention to the influence of the prism angle upon the effective aperture of the emergent beam. A model of the instrument, now in use, is described.

1. INTRODUCTION

THE need sometimes arises for an optical path much longer than the greatest dimension of the region through which the path may practicably be made to extend. Although uses in emission are not inconceivable, usually the need is for an absorption path, for instance, in the study of absorption in gases at low pressure, or of very weak absorption (i.e., very low transition probability), or of the absorbing properties of rare materials or of particular samples.

In particular, we were led to the design of the instrument described here by the desire to study the fine structure of certain atomic spectra.[1] This structure would have been spoiled by the Doppler and resonance broadening of the lines except at low pressures. At these low pressures it was estimated that paths of several tens of meters would be necessary to produce sufficient absorption for our studies. The maintenance of optical rigidity and the control of the temperature within tubes of this order of length, as well as the geometric limitations of the laboratory, presented formidable problems, which were avoided by the use of this instrument.

Several other devices for obtaining long optical paths in restricted spaces have been described;[2–7] the list of references given here is not supposed to be complete.

Briefly stated, the long-path device consists of a concave spherical mirror and, near its center of curvature, a truncated 45°–45°–90° prism, Figs. 1, 2. (Strictly, the prism angle is not quite 90°, but the question of the optimum angle is discussed separately in Section 3.) The axis of

[1] H. R. Kratz, Phys. Rev. (in preparation).

[2] L. R. Ingersoll, private communication. Mirrors at the ends of the tube.

[3] R. M. Badger, L. G. Bonner, and P. C. Cross, J. Opt. Soc. Am. **25**, 355 (1935). The tube length equals the path length, but a great relative aperture is attained with small cross-sectional area by reflections inside a polished cylindrical metal tube.

[4] F. T. Rogers, J. Opt. Soc. Am. **29**, 479 (1939). An individually adjusted mirror at the end of the tube, for each traversal.

[5] H. Beutler, private communication. A spherical mirror at one end, and at the other end an individually adjusted totally reflecting prism containing a set of images like one of the sets S, S', S'' in our Fig. 1, for each double traversal. Our solution of the problem gained impetus from a consideration of Beutler's instrument.

[6] H. D. Smith and J. K. Marshall, J. Opt. Soc. Am. **30**, 338 (1940). Optically identical with our scheme except for the use of two ordinary mirrors instead of our totally reflecting prism.

[7] J. U. White, J. Opt. Soc. Am. **31**, 756A (1941); **32**, 285 (1942). Uses three mirrors and no prism; possess the property of universal adjustment for number of traversals, similar to ours. White lists nine advantages of this ingenious triple-mirror design over ours. Number 1, conservation of aperture, is a real advantage, and probably would be important for a very large number of traversals where our aperture might vanish; yet in the production of an image upon a spectrograph slit by means of a lens, the convergence of the beam can be increased as much as desired, at the expense of image size (line length), which is usually of no importance. The later items of number 4 constitute a real advantage, but with regard to the question of brightness, numbers 4 and 5, the triple mirror device, with a brightness approximately $[\frac{1}{4}R_\lambda^{\frac{1}{2}}\mu^{-1}(\mu+1)^2]^{N+2}$ as great as ours, cf. our Eq. (7), is at a disadvantage for short wave-lengths; see our item (d), below. Regarding number 7, efficiency of use of the volume, the necessity for two mirrors side by side makes the three-mirror design only half as efficient as ours. Number 2, 3, 6, 8, and 9, or analogous statements, can be made about both designs. See note added in proof at the end of this paper.

the mirror lies in the plane that bisects the prism normal to the prism faces. A source S_0, close to the prism, is imaged by the mirror at $S_2, S_4, \cdots S_N$ close to itself after 2, 4, $\cdots N$ traversals of the space between the prism and the mirror, where N is an even integer representing the total number of times the light traverses the instrument. By two total reflections in the prism, S_2 is brought to S_2'', S_4 to S_4'', $\cdots S_{N-2}$ to S_{N-2}''. But near S_N the beam undergoes only one reflection, so that after the Nth traversal an image S_N' is formed near the center of the large prism face and the beam leaves the prism at right angles to its original direction, through a second truncation.

Some of the properties of our long-path device[8] are:

(a) The number of traversals of the length of the tube is in principle unlimited,[9] but the angular spread of the final beam is less than that of the first traversal, by an amount that becomes serious, however, only for a large number of traversals; cf. Fig. 3 and Eqs. (9) and (11), below.

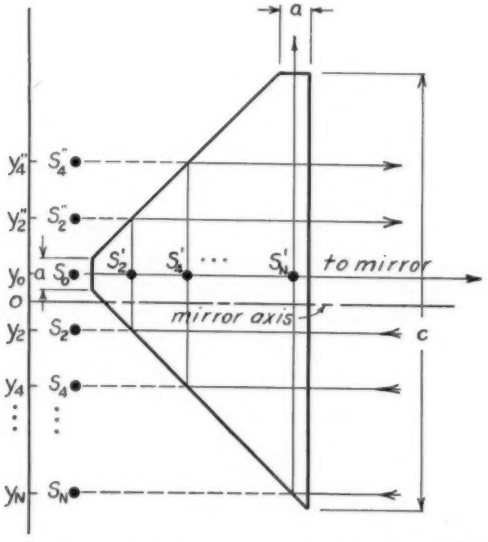


FIG. 1. The prism, and the images S_k, S_k', S_k''. When the prism is used for N^*, i.e., for the greatest value of N for which the prism was designed, the mirror axis touches the bottom edge of the first truncation, Eqs. (6).

(b) Light is conserved by image formation at alternate traversals.

(c) The image formation is without aberration, except for the aberration arising in the imaging of a small object near the center of curvature of a spherical mirror.

(d) The final image is as large, and in favorable cases almost as bright, Eq. (7), as the source. The light is reflected by an aluminum mirror at only one end of the tube, so the reflection losses are more or less than when the light is reflected at both ends by mirrors according as the reflectance of the mirror is greater or less than about 92 percent. Thus the design yields greater brightness than White's[7] in the ultraviolet, slightly less in the visible if the mirrors are silvered, and greater if aluminized, and in most instances less in the infra-red.

(e) The device is especially suitable for use with line-shaped sources such as capillary arcs, which are excellent continuous sources.

(f) The optical adjustments are simple, and their simplicity is independent of the number of traversals required.

(g) The space inside the chamber is used with almost ideal efficiency, for absorption within a cylindrical body of absorbing material.

2. GENERAL OPTICS

The problem may be considered, except for minor details, and is considered below, as a two-dimensional one of relationships in the plane of the mirror axis normal to the prism faces. Several statements that will be obvious from a consideration of elementary geometrical optics are made without proof.

The mirror can be completely described for our present purposes by:

$b \equiv$ diameter of the mirror face,
$r \equiv$ radius of curvature of the spherical surface.

The description of the prism (Fig. 1) is considerably more complicated. The quantities that must be specified accurately are:

[8] H. R. Kratz and J. E. Mack, Phys. Rev. **57**, 1059A (1940).

[9] In our tube, designed for twelve traversals, the twenty-traversal beam (path length ~83 meters) has been seen, although at least 40 percent of the light, and more if one allows for a rounding at the edge of the truncation, must have been lost at the end of the second traversal; cf. Eq. (4b) and the discussion below it. We believe that the practical limit for general purposes may lie between twenty and forty.

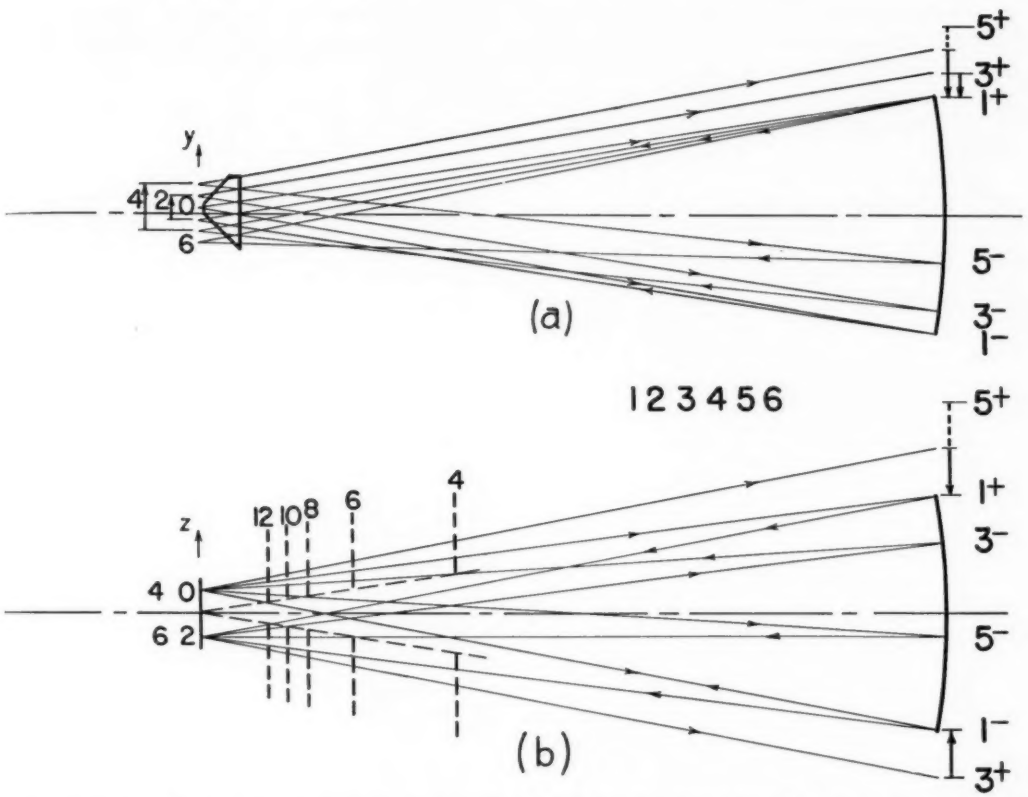


FIG. 2. Cross sections of the instrument, in the case of a prism designed and used for $N^*=6$. The spherical radius of the mirror is reduced by a factor of about 20, from the relative value actually used, Section 4. The (odd) numbers at the right show the limits of the successive beams approaching the mirror; the sign remains unchanged for each edge of the beam, e.g., ray 1^+ returns to the side or sides indicated by 3^+ and 5^+. The arrows on the right show losses by spilling off the edge of the mirror; the dashed part of the arrow at 5 shows the part of the beam previously lost at 3. The successive limiting rays can be distinguished conveniently by reference to the numbers vertically in line with the arrows on the rays. The (even) numbers at the left show the successive images of the point labeled 0.

a. The xy plane, drawn for the case $\varphi=90°$ exactly. The source may be considered as a point S_0 or 0, at the ordinate y_0. The arrows at the left show the upward shifts of $2ky_0$, Eq. (5), between the image S_k at the end of the kth transition and the effective object S_k'' at the beginning of the $(k+1)$th transition. For the case pictured here the deviation ϵ equals zero, so the length of the full line arrow (right) at the edge of the mirror at the end of the $(k+1)$th transition is also equal to $2ky_0$.

b. The xz plane. Here 0 is any point at ordinate z_0 on the linear source of length h, and $z_6=-z_4=z_2=-z_0$. The arrows at the right show that the spilling at the end of each odd-numbered transition occurs at the $-$ edge of the beam, at alternate edges of the mirror, always in the amount $2|z|$. The interrupted heavy dashed lines are the imaginary diaphragms, Eqs. (12), at the edges of the even numbered beams; diaphragms number 4 and 6 can be seen to graze the fourth and sixth beams, respectively.

$\varphi \equiv$ the value of the "right angle" of the prism (Section 3),

$a \equiv$ the width of the truncation (hereafter called the first truncation) at the "right-angle" vertex,

$c \equiv$ the width of the largest prism face = twice the distance between the largest prism face and the first truncation,

$\mu \equiv$ the refractive index of the prism.

The second truncation shown at the top of the prism in Fig. 1 is not essential to the instrument, but is useful in practice. Its width must be at least a, and for simplicity we assume it to be just a.

The exact orientations of the large face and the truncations with respect to the small faces are unimportant; but they are approximately normal, normal, and parallel, respectively, to the bisector

of the right angle. Terms of direction in the following discussion, such as "top," etc., imply the orientation shown in Fig. 1, where the truncated 90° angle is at the left, the second truncation at the top, and the mirror at the right.

The ideal distance of the 90° vertex of the prism from the center of the mirror is

$$x_0 = r + \tfrac{1}{2}(c+a)(1-\mu^{-1}), \tag{1}$$

an expression in which the second term may be recognized as the usual focus correction for a flat plate. Although the correction is of some practical importance, the distance need not be fixed with optical accuracy; a displacement of the source by Δx from this ideal distance x_0 introduces a longitudinal displacement $-N\Delta x$, with respect to the prism, in the image formed after N traversals. Strictly, one would expect the actual source to be a distance $\frac{1}{2}a\mu^{-1}$ from the truncation, so that its image S_0 by refraction in the first truncation would coincide with the 90° vertex; actually the distance is immaterial as long as the source is enough wider than a to fill the whole truncation as seen from every part of the mirror. In the remainder of the discussion it will be unnecessary to consider distances from the mirror, for all the object and image distances can be considered equal to r.

The exact orientation of the prism is unimportant, on account of the well-known angular independence of the ray-reversing property of a 90° pair of mirrors; but it is mathematically convenient to consider the bisector of the "right angle" to be parallel to the mirror axis, which we shall take as our $y=0$ axis. On the other hand, the value of N for the beam that emerges from the second truncation is exceedingly sensitive to the orientation of the mirror, or in other words to the y coordinate y_0 of the bisector of the 90° angle, which is the same as the y coordinate of the center of the source S_0.

After two traversals, the image S_2 of the center of the source, produced by the mirror, is at $y_2 = -y_0$. The two successive images produced by the totally reflecting prism are a real one S_2' on the bisector of the 90° angle and a virtual one S_2'' at $y_2'' = 3y_0$, the latter of which acts as the object for the next image formation by the mirrors. In general,

$$\left.\begin{aligned} y_k = -y_{k-2}'' &= -(2y_0 - y_{k-2}) \\ &= -(k-1)y_0 \end{aligned}\right\} K = 2, 4, \cdots N. \tag{2a}$$

$$y_k' = y_0 \tag{2b}$$

$$y_k'' = 2y_0 - y_k = (k+1)y_0 \tag{2c}$$

In order that the beam emerge through the second truncation after N traversals it is necessary that the image S_N be a distance $\frac{1}{2}c$ below the source, i.e., that

$$y_0 - y_N = \tfrac{1}{2}c \tag{3}$$

or from Eqs. (2a) and (3), that

$$y_0 = \tfrac{1}{2}cN^{-1}. \tag{4a}$$

From Eqs. (2a), (2c), and (4a)

$$y_k'' - y_k = 2ky_0 = kcN^{-1}. \tag{5}$$

These equations are subject to the condition that

$$y_0 \geqslant \tfrac{1}{2}a, \tag{4b}$$

i.e., that the mirror axis intersect the slant face of the prism rather than the truncation; for if y were less than $\frac{1}{2}a$, part of the beam would be lost through the truncation at the end of the second traversal. Equation (4a) shows that if the prism is designed so that

$$c = N^*a, \tag{6a}$$

then for the beam to emerge after N^* traversals

$$y_0 = \tfrac{1}{2}a \tag{6b}$$

and for it to emerge after any smaller even number of traversals N,

$$y_0 = \tfrac{1}{2}aN^*N^{-1}. \tag{6c}$$

Thus, if the prism is lowered at a uniform rate from $y_0 = \frac{1}{4}N^*a$ to $y_0 = \frac{1}{2}a$, beams that have traversed the prism-mirror path 2, 4, $\cdots N^*$ times will emerge from the second truncation with increasing frequency, until the (N^*-2)th and N^*th beams come out practically adjacent.

The relationship of the brightness B_N of the beam emerging after N traversals, to the bright-

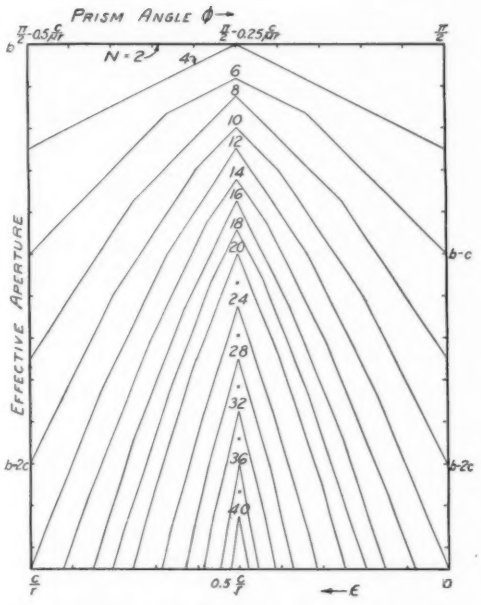


FIG. 3. Effective width of the mirror in the plane of Fig. 1, as a function of φ or ϵ for any even number of traversals. The effective width, or aperture$_y$, is expressed in terms of the actual mirror width b and the prism width c, Eq. (10a). The abscissa is expressed at the top in terms of the prism angle φ, and at the bottom in terms of the angle ϵ between successive rays at the prism, Eq. (8). The figure is symmetric about a vertical line through the center. If the curves were carried beyond the side edges of the figure they would be found to extend indefinitely in both directions with the slopes shown just inside the edges, i.e., pointing toward the top center point of the figure.

ness B_0 of the source, is[10]

$$B_N = B_0 R_\lambda{}^{\frac{1}{2}N}[4\mu(\mu+1)^{-2}]^{N+2} l_\lambda{}^N, \qquad (7)$$

where

$R_\lambda \equiv$ the normal reflectance of the mirror for light of the wave-length considered,
$\mu \equiv$ the refractive index of the prism,
$l_\lambda \equiv$ the fraction of the incident light transmitted by the medium in a single traversal.

[10] In Eq. (7) the refractive property of the medium has been neglected. Some of the light is refracted out of the beam in the neighborhood of an absorption line by anomalous dispersion, and thus a slight false broadening of the line is effected. For a line in the near ultraviolet with a Ladenburg f value of 1, in the case of a gas at 1-mm pressure, and an instrumental resolving power of 10^5, the effect upon the intensity at the edge of a line would be several percent; cf. S. A. Korff and G. Breit, Rev. Mod. Phys. **4**, 471 (1932), especially Eq. (37) and tables. In our work (reference 1) the effect is certainly negligible.

It will be shown in the next section that for large values of N the angular divergence of the light from the final image S_N' in the prism is considerably less than br^{-1}, both in the plane of Fig. 1, cf. Eqs. (10), and in the plane perpendicular to the figure, cf. Eq. (11) or Eqs. (12).

3. CALCULATION OF THE EFFECTIVE APERTURE OF THE MIRROR

In this section we shall consider the limitations imposed by the mirror upon the divergence of a beam that has executed N traversals before emerging. We shall consider first the plane of Fig. 1 (xy plane), and then in the next to last paragraph of this section, the projection in the plane $y=$constant (xz plane).

In the xy plane, Fig. 2a, it is sufficient to consider a point source S_0 at the ordinate y_0. The effective aperture of the mirror after two traversals is evidently b, regardless of the prism angle, since the beam is limited only by the mirror itself. In the case of a prism angle exactly $\frac{1}{2}\pi$, which has the familiar property of returning each ray parallel to itself, after the second traversal the whole beam is shifted upward, according to Eq. (5), a distance $4y_0$ or $2cN^{-1}$, and the effective aperture of the mirror at the end of the fourth traversal is only $b-2cN^{-1}$. More generally, for a prism angle exactly $\frac{1}{2}\pi$, the effective aperture of the mirror after N traversals is reduced by the amount

$$\sum_{\substack{\text{even}\\ k=0}}^{N-2} kcN^{-1}, \quad \text{or} \quad \tfrac{1}{2}c(\tfrac{1}{2}N-1)$$

by the escape of light past the top of the mirror.

A slight reduction of the prism angle alters the direction of the third and later beams to bring back upon the mirror some or all of the light that would have passed over the top of the mirror in the case of a prism angle exactly $\frac{1}{2}\pi$, thus increasing the effective aperture of the mirror. Complete compensation in the effective aperture can be made for the $N=4$ beam by modifying the prism angle so as to direct the beam downward by an angle $\frac{1}{2}cr^{-1}$. But for higher values of N, the greatest compensation possible is incomplete,[11] because the magnitude of the

[11] Unless one were to shape one or both of the prism faces especially—a possibility that we shall not consider in detail.

image shift increases with increasing k, cf. Fig. 2a and Eq. (5), and only a single variable, *viz.*, the prism angle φ, is at one's disposal to compensate for all the different shifts. The angle ϵ between the forward direction of the $(k+1)$th beam after it leaves the prism and the backward direction of the kth beam before it enters the prism has the small value

$$\epsilon = 2\mu(\tfrac{1}{2}\pi - \varphi). \tag{8}$$

The general problem of the dependence of the effective aperture of the mirror upon the prism angle, for all even values of N, is an elementary one, but it is complicated by the fact that the shifting of successive beams off the edge of the mirror introduces a broken-line dependence of the effective aperture of the mirror upon the prism angle.

The solution is shown graphically[12] in Fig. 3. The calculation shows that the dependence of the effective aperture upon ϵ is symmetric about the optimum value

$$\epsilon_{\text{optimum}} = \tfrac{1}{2}cr^{-1} \tag{9}$$

which, it will be noticed, is independent of N. The relationship for $\epsilon \leqslant \frac{1}{2}cr^{-1}$ is:

$$\text{aperture}_y = b - \tfrac{1}{2}c(\tfrac{1}{2}N-1) + \epsilon r(\tfrac{1}{2}N-1-l) + (l^2+l)cN^{-1}, \quad \epsilon \leqslant \tfrac{1}{2}cr^{-1} \tag{10a}$$

where b, c, N, and r are defined early in Section 2, ϵ is defined by Eq. (8), and $l \equiv$ the greatest integer in the number $\frac{1}{2}\epsilon Nrc^{-1}$ (but for $\epsilon < 0$, $l \equiv 0$).

[12] The following is a simple interpretation of Fig. 3. The half of each curve that slopes upward with increasing ϵ represents the condition in which, with increasing ϵ, light is collected that with smaller ϵ would have spilled over the top of the mirror. It will be noticed that in the neighborhood of $\epsilon = 0$ the slope of the line is such that, extrapolated, it would represent an aperture equal to b at $\epsilon_{\text{optimum}}$. This initial rate of gain with increasing ϵ is not maintained, however, for eventually the $k = 3, 5, \cdots$ beams successively spill over the bottom edge. The rate of spilling of the kth beam is proportional to k. Each break represents a new beam (new value of k) slipping off the bottom, but the net effect is to increase the aperture until the top of the $(N-1)$th beam comes into the mirror, at $\epsilon_{\text{optimum}}$. After that an increase in ϵ only spills successive new beams off the bottom until the $(N-1)$th slips off at the rate at which it came on. The curve is broken into $\frac{1}{2}N+1$ or $\frac{1}{2}N$ segments according as $\frac{1}{2}N$ is odd or even (except for the trivial case $N=2$). The difference arises from the fact that when $\frac{1}{2}N$ is odd the (optimum) value of ϵ at which the top of the mirror ceases to become important with increasing ϵ occurs at the middle of the range in which the $\frac{1}{2}N$th beam is slipping off, and when $\frac{1}{2}N$ is even this optimum occurs just between the slipping off of the $(\frac{1}{2}N-1)$th and that of the $(\frac{1}{2}N+1)$th transition.

The relationship for $\epsilon > \frac{1}{2}cr^{-1}$ can be found by the symmetry about $\frac{1}{2}cr^{-1}$.

For any value of N the optimum aperture is

$$\text{optimum aperture}_y = b - \tfrac{1}{4}c[\tfrac{1}{4}N - 1 + \delta(\tfrac{1}{2}N, \text{odd})N^{-1}] \tag{10b}$$

and in the immediate neighborhood of the optimum

$$\frac{\partial(\text{aperture}_y)}{\partial|\epsilon - \epsilon_{\text{optimum}}|} = -\tfrac{1}{2}r[\tfrac{1}{2}N - \delta(\tfrac{1}{2}N, \text{odd})], \tag{10c}$$

where the symbol $\delta(\frac{1}{2}N, \text{odd}) \equiv \begin{cases} 1 \text{ if } \frac{1}{2}N \text{ is odd.} \\ 0 \text{ if } \frac{1}{2}N \text{ is even.} \end{cases}$

In the xz projection, Fig. 2b, the situation is much simpler. If the source has a length h it may be considered as a short line of length h, normally bisected by the x axis. For simplicity the prism may be considered as replaced by a plane mirror coincident with the source, for the reflections in the prism do not influence the inclination of the beam in the z direction. We are interested in the effective aperture of the mirror for the various parts of the source, i.e., for a point S_0 at an ordinate which we may call z_0 or z, where $|z| \leqslant \frac{1}{2}h$. Elementary considerations show that the images occur alternately at the two positions $z_N = (-1)^{\frac{1}{2}N}z_0$, where N is an even integer as before,[13] and that portions of the beam of width $2|z|$ are spilled at each odd-numbered traversal after the first. Therefore,[11] the effective aperture of the mirror in the xz plane after N transitions extends for positive z, between $(-1)^{\frac{1}{2}N} \cdot \frac{1}{2}b$ and $(-1)^{\frac{1}{2}N-1}[\frac{1}{2}b - (N-2)z]$; for negative z the signs of both expressions must be changed. In other words the effective aperture, for every point of the source except the center, is an asymmetric one with a width at the mirror limited to

$$\text{aperture}_z = b - (N-2)|z|. \tag{11}$$

This limitation can be represented, independent of z, by an imaginary diaphragm with parallel edges at distances

$$\pm\text{stop half-width} = \pm\tfrac{1}{2}b(N-1)^{-1} \tag{12a}$$

from the $z=0$ axis, lying in the plane that cuts the axis normally, ahead of the image S_N by

[13] It is not necessary here, as it was in the case of the xy plane, to distinguish between k and N.

the distance

$$\text{stop-to-image distance} = r(N-1)^{-1} \qquad (12b)$$

in the path of the Nth traversal.

The different limitations upon the effective width of the mirror in the y and z directions, which have been discussed in this section, sometimes need to be borne in mind in the consideration of such questions as the choice of condensing lens and of lens and slit distances, when one uses the emergent beam to illuminate the slit of a spectrograph.

4. EXPERIMENTAL

The instrument that we have built is a vacuum tight chamber 423 cm long, which consists mainly of a 3-inch O.D. steel tube separated into three segments by copper bellows about 30 cm from each end. Each of the two end segments is surrounded by a concentric tube extending beyond it to allow space for the installation of optical parts. The total inside volume of the chamber is a little less than 20 liters. The space between the two concentric tubes serves as a water jacket.

The end segments are silver-soldered to heavy brackets, which in turn are longitudinally slotted and bolted near the ends of a $2\frac{1}{2}$ by 4 inch steel I-beam, which thus holds the optical parts rigidly in adjustment. The slots offer all necessary adjustability in mirror-to-prism distance, Eq. (1).

Rigidly attached to each end of the long center segment of the tube is a thin cylindrical steel sleeve, somewhat longer than the bellows, so that it fits into the end segments. The radial clearance between these sleeves and the end segments is only about 0.2 mm. The sleeves serve the double purpose of guiding the center segment into alignment with the ends and keeping the bellows cleaner than they would have been without this internal protection.

At the mirror end of the tube is a plate, fixed in the tube, and mounted against it, an adjustable plate to which the mirror is fastened. The mounting is a three-point mounting with the points arranged in the form of an isosceles right triangle, with vertical and horizontal sides. The mounting post at the right angle is fixed, and the other two have screw adjustments. The screws can be adjusted from the outside by means of auxiliary 1/8-inch rods through the end plate. These and all the other rods through the vacuum walls are passed through by means of Wilson seals.[14] Of these two mirror adjustments, the one that moves the image horizontally is necessary. The other is redundant (cf. the vertical motion of the prism, described below) but convenient for preliminary alignment. The mirror is made of glass, aluminized on the front surface. Its diameter, b, is 7.3 cm, and its radius of curvature r 416 cm.

At the prism end of the tube there is a quartz window 3.8 cm in diameter at the center of the end plate, and another of the same size on the top of the cylindrical surface of the tube, above the second truncation of the prism. (Two auxiliary totally reflecting prisms, the first mounted directly above the top window, have been found convenient for adjusting the beam without the necessity of moving the whole instrument.) The source is a General Electric Type H-6 water-cooled mercury arc mounted horizontally in its regular quartz water jacket, which in turn is placed inside a water-cooled brass cylinder 7.6 cm long and 3.8 cm in outside diameter that allows light to pass only in the desired general direction through a slot. The source assembly is almost in contact with the window of the end plate. It has a vertical screw adjustment. Below the window on the end plate is a seal for a rod that reaches half way down the tube, inside, and may be used as a hammer to break a glass capsule containing the material to be studied. The rod is built in sections. It may be almost entirely removed, and the remaining part covered with a plate, almost flush with the end plate, that offers no obstacle to the motion of the lamp.

The prism is mounted in the orientation shown in Fig. 1 on a vertical rack, with two external adjustments. The rack itself can be moved about a vertical axis, and the prism can be moved vertically along the rack. The latter motion is the one that is finally used to determine the number of traversals. The amount of this adjustment required to allow the use of all possible values of N from 2 to N^* is, according to Eq. (4a) $\frac{1}{2}c(\frac{1}{2}-1/N^*)$. The effect of this adjustment is described just below Eq. (6c). The actual

[14] R. R. Wilson, Rev. Sci. Inst. **12**, 91 (1941).

prism is of crystal quartz, with the optic axis normal to the plane of Fig. 1. The prism was designed[15] for a maximum number of traversals, N^*, equal to 12; cf. Eq. (6a). The width a of the first truncation, is 2 mm, and the width of the prism c is 24 mm. The length h of the prism normal to the plane of Fig. 1 is 25 mm. The optimum prism angle $\varphi_{\text{optimum}}=\frac{1}{2}\pi-\frac{1}{4}c(r\mu)^{-1}$, cf. Eqs. (8), (9), would have been 89° 56′ 47″, which for $N=12$ would have yielded an aperture, Eq. (10b), of $b-\frac{1}{2}c$ or 61 mm. The actual value of the angle of the prism is 89° 57′ 40″, yielding, for $N=12$, an aperture, of $b-0.7c$ or 57 mm, as may be seen from Eq. (10a) or Fig. 3. The virtual stop before S_{12}, Eqs. (12), is $b/11$ or 6.6 mm wide, and $r/11$ or 38 cm before S_{12}.

Almost the entire length of the center section of the tube is heated by means of a 110-volt d.c. current passing through 40 meters of 0.6 by 6.3 mm nickel ribbon. The nickel ribbon is wound over a covering of Micanite for electric insulation, in a spiral with a pitch of 2.5 cm, except for the last 5 cm at either end, where the pitch is halved to compensate for the calculated heat losses to the water jackets. In order to reduce these heat losses and maintain a uniform temperature the entire length of the tube, the steel tube between the ends of the nickel windings and the copper bellows is turned down to a thickness of 0.5 mm for a distance of 5 cm. The entire center section of the tube is surrounded with a three-inch layer of magnesia steam-pipe insulation.

The temperature of the tube is maintained constant to within ±0.1°C by means of a balanced Wheatstone bridge circuit.[16] The nickel windings on the tube serve as one arm of the bridge; the other three arms are of Advance wire wound on a water-cooled cylinder. The temperature coefficient of resistance of the nickel is high, whereas that of the Advance wire is negligible. Thus, when the temperature of the nickel ribbon changes, it changes the balance of the bridge. A Weston model 534 sensitive relay takes the place of the galvanometer, this relay controlling less sensitive relays, which in turn control the heating current. Two arms of the bridge are adjustable so that the bridge circuit can be set for any temperature between 100° and 500°C.

Note added in proof: Since this paper was written, Mr. John U. White, to whom we submitted our footnote 7 for criticism, has kindly called attention to the following two facts: With regard to his listed advantage number 1, image size is of advantage in monochromators designed for the purpose of emitting a large amount of flux, although it is not of advantage in photographic instruments. Similarly, with regard to his listed advantage number 7, if images or large separations between images are desired, the three-mirror design may be preferable because by forming only one image per double traversal instead of two on the line $S_4''\cdots S_{N-1}$ of Fig. 1 it needs only half as much space at this end of the absorption cell as the prism design.

[15] In the light of our present experience (reference 9) we recommend a considerably smaller value of a and a greater consequent N^*. Regarding the value of φ, it may be more economical to order several prisms ground to 90° with ordinary accuracy and select the one with the best φ than to order one especially ground to the required accuracy.

[16] H. S. Roberts, J. Opt. Soc. Am. **6**, 965 (1922).

Attributive Limens in Selected Regions of the Munsell Color Solid

BLANCHE R. BELLAMY, *Munsell Color Company, Baltimore, Maryland*

AND

SIDNEY M. NEWHALL, *The Johns Hopkins University, Baltimore, Maryland*

(Received May 25, 1942)

INTRODUCTION

A COLOR solid represents different colors by different locations, and color differences by spatial differences. Each Munsell color is designed as a unique combination of hue, value, and chroma which can be represented by a particular spatial point. Hue, value, and chroma are the Munsell analogues of psychological hue, lightness, and saturation. An attributive difference limen is the amount of difference, in some dimension, required to permit correct perception of difference in a prescribed proportion of many trials.

In the present study, difference limens for hue, value, and chroma were separately secured at a variety of locations spotted throughout the volume of the Munsell solid.[6] During a period of two years, three experienced observers have furnished data for approximately 500 such color-difference limens.* On the basis of these results, mean determinations have been computed for 41 representative regions of the solid. Figure 1 indicates the distribution of the points at which these liminal determinations were made. The various levels of constant value are shown and the particular samples from which the liminal departures were measured are indicated by black circles in white spaces. Levels 2/, 5/, and 8/ are represented at the right side of the figure and levels 3/, 4/, and 6/ at the left. A relatively large number of limens was taken at the 5/-value level because that is considered to be the most representative. A number of limens were measured at levels 2/ and 8/, as well as 5/, because these three levels would represent dark, medium, and light. A few measurements were made at intermediate value levels to secure instances of strong chroma and to provide more complete chroma sequences on the so-called home value levels of particular hues. The 7/-value chart is not shown in the figure, but a threshold for the corresponding achromatic sample was secured at that level to complete results for the neutral series.

Tables I and II exhibit the mean liminal determinations in detail, together with the respective ranges of variation. In both of these tables, the first column contains the Munsell notations corresponding to the designated locations in Fig. 1.

Though limited in scope, these appear to be the only systematic data available which provide comparative threshold measurements corresponding to all three fundamental dimensions. It was believed that the detailed results would be of interest in relation to such important problems as the setting of perceptual tolerances,[13] the respacing of the Munsell colors,[8] the equation of the attributive scales,[10] and the equality of just noticeable differences (reference 3, p. 143). This study was mainly motivated, however, by the simple desire to secure dependable information with respect to the psychological magnitude of the Munsell scale units; that is, to discover the size of the units as measured in terms of the difference limen.

PROCEDURE

General Method

The method of making the liminal measurements was briefly as follows: The observer looked into the open end of an observation box and saw a bipartite field whose dividing line was a vertical arc formed by a pair of partially overlapping color disks. These Maxwell disks spun so closely behind an opening in a masking screen that the latter could serve as the surround of the bipartite field which was itself perceived as surface. By adjusting the amount of overlap of the disks, the halves of the field could be made to differ by small and known amounts with respect to the attribute under test. By a series of observations during which the difference between the

* The observers were B. R. Bellamy, A. E. O. Munsell, and S. M. Newhall.

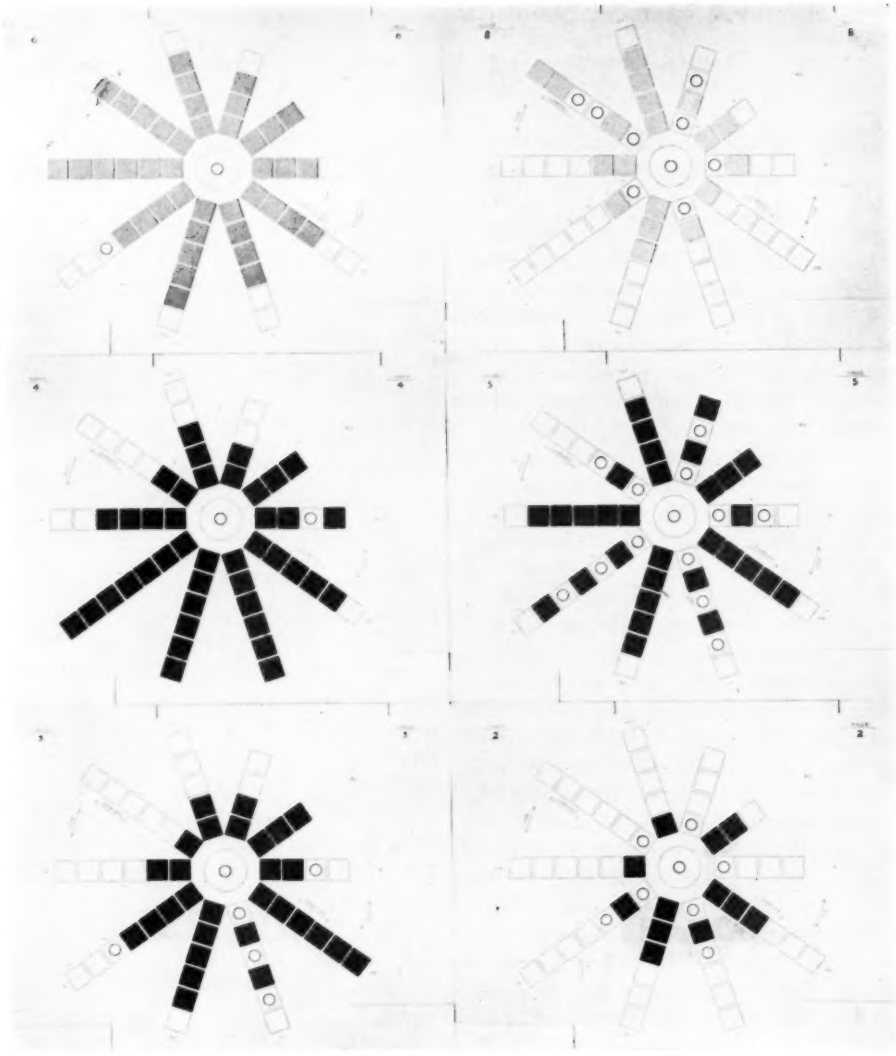


FIG. 1. Circles show locations in the Munsell solid of the common origins of difference limens of hue, value, and chroma.

halves of the field was repeatedly changed and the observer reported the corresponding color differences perceived, data could be secured which revealed perceptibility as a function of difference. The difference which would yield liminal perceptibility then could be computed by interpolation, in accordance with the accepted definition.

The difference threshold is usually defined as that amount of difference required to yield correct perception of the direction of difference in 50 percent of numerous independent trials under favorable conditions of viewing (reference 3, p. 112). The experimentation was by a variation of the method of constant stimulus differences (reference 3, p. 186) which is known as the method of group presentation,[11] while the calculation was by an abbreviated form of the con-

stant process.[7] Originally, the limens were computed in terms of percent disk sectors exposed, but subsequently they were transformed to the 1929 Munsell notation,[4] by application of the standard formulae for hue and chroma[9] and the Glenn-Killian data for value.[2] The limens can be re-expressed in terms of the standard I.C.I. colorimetric coordinate system after an adequate transformation has been prepared for the purpose.

Conditions of Observation

The size of a limen varies with the conditions of observation, and in general, the more favorable the conditions the smaller the limen. The conditions of the present experiment were considered to be very favorable. The observers all knew its purpose and they were practiced in distinguishing the attributive dimensions. On the other hand, any special bias of expectation was avoided by the fact that the experimenter was always the one to adjust the disks and the observer never saw or was told what sectorial adjustment had been made. Only two disks were ever used at one time; the smaller disk partially overlapped the larger which served as standard. The periphery of the smaller disk provided the dividing line of the field. Disks were chosen which were designed to differ only in regard to the attribute under test, another condition which favored fine discrimination.

The observer made his judgment by comparing the mixture half of the field with the standard half; and he was instructed to make his judgments in the psychophysical tradition, i.e., rather rapidly to avoid vacillation and with moderate confidence to avoid guessing. The terms of judgment were, in the case of chroma "weak," "strong," "match," in the case of value "dark," "light," "match," and in the case of hue "R," "Y," "G," "B," "P," "match." The term "match" as thus used judgmentally, referred only to the attribute in question. Frequently, the other two attributes refused to remain perceptually constant so it was necessary to caution the observer to restrict intentionally his judgment to the particular dimension in question. In addition, the observer was cautioned not to depend upon disappearance of the dividing line as the criterion of "match." As a matter of fact, the dividing line often disappears and so the observer was instructed to base his judgment on the relative appearance of the two areas themselves. If contrast difficulties were disturbing, he was asked to confine his comparison to the more central portions of the areas.

The observers were on guard also against another difficulty which has long been recognized by psychophysicists and which is most likely to be bothersome when observing near the threshold. This is the difficulty of attributive confusion. Brown and Hoisington, for instance, had noticed that weak chromatic colors are often incorrectly named and that a film experience slightly lighter or darker than its surroundings is likely to arouse an illusory experience of hue.[1]

Various external conditions were held constant throughout the observations. The observer was light-adapted and viewed the bipartite field binocularly from a distance of 24 inches. This central field subtended a visual angle of 6.5° while surrounding it was the neutral screen or ground of an angular diameter of nearly 30°. The illuminance of both central field and surround was 40 ft.-c from a common source of Macbeth daylight at 5740°K; and the surround always closely matched in lightness the standard half of the central field. A number of interchangeable surround surfaces of assorted reflectances were required to maintain this lightness match. Beyond the limits of the 30° surround, the visual field was filled out by the mat neutral interior walls of the observation box which tapered toward the field end. Between observations the observer looked toward the middle of a mat neutral wall with an illuminance of about 40 ft.-c and a reflectance of 0.2.

Meticulous care was required to avoid disturbing errors from defects in the Maxwell disks. Minor imperfections which would be tolerated without question in all ordinary uses must be scrutinized when very small differences are themselves the object of inquiry. For instance, the scale which was employed for measuring the size of the comparison sector during adjustment would leave slight arc-shaped marks on the disk surface unless handled with caution. During rotation such marks could produce a slightly altered specification. The value thresholds at low value levels provided a somewhat different prob-

TABLE I. Attributive limens from chromatic origins.

Common origin (1929 notation) *H v/c*	Hue Mean	Hue Range	Value Mean	Value Range	Chroma Mean	Chroma Range
R 8/2	0.690	0.310–1.230	0.038	0.024–0.049	0.144	0.124–0.164
R 6/10	0.590	0.280–0.850	.037	.033– .040	.510	.440– .570
R 5/10	0.560	0.430–0.800	.040	.036– .046	.340	.250– .480
R 5/6	0.480	0.290–0.730	.027	.025– .030	.280	.204– .342
R 5/2	0.580	0.300–0.800	.035	.031– .038	.084	.082– .085
R 3/10	0.410	0.240–0.510	.011	.009– .015	.350	.250– .450
R 2/6	0.180	0.050–0.370	.013	.004– .017	.220	.198– .240
R 2/2	0.690	0.640–0.750	.008	.004– .014	.065	.056– .070
Y 8/8	0.670	0.600–0.740	.043	.041– .045	.312	.224– .408
Y 8/6	0.327	0.100–0.460	.026	.022– .032	.196	.138– .306
Y 8/2	0.410	0.260–0.550	.025	.023– .029	.102	.070– .150
Y 5/6	0.280	0.250–0.340	.010	.002– .021	.300	.240– .330
Y 5/2	0.700	0.680–0.750	.028	.022– .038	.132	.122– .144
Y 2/2	0.250	0.150–0.310	.006	.002– .010	.051	.034– .072
G 8/6	0.470	0.350–0.600	.024	.018– .029	.370	.342– .408
G 8/2	0.533	0.340–0.700	.029	.022– .038	.112	.074– .166
G 5/6	0.390	0.290–0.470	.003	.012– .008	.339	.255– .390
G 5/2	0.590	0.490–0.750	.032	.024– .039	.158	.130– .187
G 2/2	0.340	0.230–0.540	.004	.003– .006	.055	.034– .070
B 8/2	0.850	0.590–1.070	.028	.023– .031	.091	.072– .124
B 5/6	1.040	0.980–1.150	.030	.023– .036	.300	.250– .380
B 5/2	0.940	0.800–1.130	.026	.018– .031	.150	.070– .190
B 4/6	0.627	0.500–0.700	.031	.015– .030	.230	.108– .312
B 3/6	0.750	0.290–1.160	.008	.003– .013	.322	.294– .360
B 2/2	0.850	0.550–1.400	.008	.003– .012	.082	.068– .096
P 8/2	0.468	0.330–0.625	.019	.014– .024	.147	.136– .166
P 5/10	0.253	0.090–0.350	.029	.025– .032	.787	.680– .850
P 5/6	0.230	0.160–0.280	.027	.025– .028	.466	.420– .510
P 5/2	0.413	0.330–0.460	.033	.030– .038	.146	.130– .154
P 3/10	0.200	0.160–0.260	.003	.002– .006	.560	.460– .670
P 3/6	0.320	0.260–0.420	.021	.015– .028	.416	.270– .540
P 3/2	0.600	0.500–0.720	.021	.019– .022	.201	.154– .270
P 2/6	0.190	0.175–0.220	.011	.004– .018	.360	.192– .540
P 2/2	0.272	0.175–0.380	.006	.003– .011	.113	.056– 210

TABLE II. Attributive limens from achromatic origins.

Common origin (1929 notation) *Hv/*	Value Up Mean	Value Up Range	Value Down Mean	Value Down Range	Red Mean	Red Range	Yellow Mean	Yellow Range	Green Mean	Green Range	Blue Mean	Blue Range	Purple Mean	Purple Range
N8/	0.029	0.023–0.036	0.027	0.025–0.029	0.101	0.086–0.126	0.155	0.114–0.186	0.120	0.104–0.130	0.091	0.060–0.108	0.113	0.074–0.150
N7/	.019	.012– .029	.021	.014– .034	.082	.060– .116	.104	.056– .164	.089	.052– .150	.087	.050– .150	.167	.114– .210
N6/	.015	.002– .036	.030	.018—.038	.083	.056– .128	.091	.032– .166	.096	.080– .124	.080	.038– .126	.123	.096– .150
N5/	.012	.008– .014	.023	.021– .026	.088	.056– .126	.098	.076– .128	.089	.044– .126	.190	.090– .130	.137	.100– .174
N4/	.015	.008– .026	.020	.014– .028	.067	.044– .094	.072	.050– .086	.089	.086– .096	.074	.036– .116	.109	.060– .150
N3/	.005	.002– .008	.021	.015– .028	.018	.005– .032	.087	.072– .096	.072	.056– .090	.053	.026– .104	.079	.044– .110
N2/	.012	.008– .014	.013	.013– .013	.048	.038– .062	.047	.034– .056	.054	.036– .074	.053	.046– .064	.077	.044– .100

lem. These thresholds are themselves exceptionally low, as might be expected on the basis of Weber's law. Consequently, very slight alterations in the reflectivity of the papers could cause concern. If the radial slit in the comparison disk was not cut perfectly straight and clean, or if a little paint chipped from the paper along the radial cut, light specks would be produced, and they could summate and merge when whirling. Tests demonstrated that defects of either type could affect the threshold determinations. As a result, disks were frequently renewed or remade during the course of the experimentation.

Influence of Adaptation

During observations the authors have repeatedly noticed that the influence of duration of fixation on the stimulus field is different for

saturation than for lightness. The chromatic adaptation accompanying fixation appears to reduce the size of a perceived chroma difference, or saturation difference, between the two halves of the field. In the case of value the reverse is the case, a fact which agrees with usual experience; some little time is required to bring out the maximum apparent difference in the lightness of the two areas compared. The reversed effect of adaptation was most apparent with weak colors and small differences. We made some effort to employ optimum times of fixation for the different dimensions but this amounted to little more than adopting a practice of adapting for a shorter time to chroma than to value. Incidentally, this effect merits independent investigation for it suggests a fundamental difference in the two quantitative dimensions of color sensation.

ATTRIBUTIVE LIMENS

Principal Data

Table I presents the mean limens for hue, value, and chroma which were taken at each of the 34 chromatic origins indicated in the first column. Each mean is followed by the range of the limens of the individual observers from which the mean was computed. The various entries are in terms of the respective Munsell attributive steps or units. The first row of entries in the table shows, for instance, that at the location R 8/2 in the color solid, the mean hue limen was 0.690 Munsell hue step; the mean value limen was 0.038 value step; and the mean chroma limen was 0.144 chroma step.

Table II presents similar mean limens for R, Y, G, B, and P, as taken at the seven achromatic origins indicated in the left column. In reality these are not difference limens but the absolute chroma thresholds for the several hues in question; they show how much chroma was required to yield a hueful perception. The first column of entries in this table includes difference limens for value at the same seven achromatic origins. The term "up" at the head of this first column indicates that those limens represent just perceptible increases above the origin whereas the "down" column refers to just perceptible decreases below the origin.

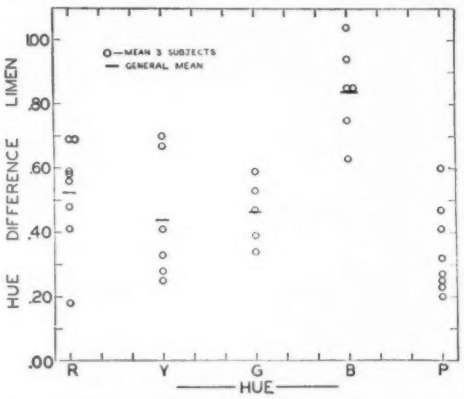


FIG. 2. Difference limen for hue as a function of hue.

The reliability of the data is roughly indicated by the ranges of the individual differences in the limens, as shown in Tables I and II. This variability is inevitably large partly because of the inherent difficulty of threshold judgments and partly because these ranges include both the intra- and the inter-individual variation. In order to secure a more definitive estimate of the precision of measurement, a sample of about 60 determinations was repeated, by two of the observers, on a different date. Table III exhibits a comparison of the repeat with the original

TABLE III. Percent departure of repeat limen from original limen.

Common origin (1929 notation) *H v/c*	Hue		Value		Chroma	
	Mean	Range	Mean	Range	Mean	Range
R 8/2	115	19–210	76	75– 76	11	3–19
R 5/10	43	27– 58	11		30	28– 32
R 2/2	61	59– 62	36	29– 43	96	71–120
Y 8/2	30	24– 35	18	9– 26	10	6– 14
Y 5/6	32	16– 47	97	50–143	15	6– 24
Y 2/2	14	7– 20	15	0– 29	53	
G 8/2	82	70– 94	53	41– 64	19	14– 23
G 5/6	22	4– 40	220	140–300	16	4– 28
G 2/2	54	19– 88	75	50–100	43	20– 65
B 8/2	30	28– 31	26	13– 38	75	
B 5/6	25	8– 42	15	13– 17	10	1– 18
B 4/6	28	26– 29	64	60– 67		
B 2/2	26	15– 36	33	0– 66	16	9– 23
P 8/2	57	40– 73	6	0– 11	11	6– 15
P 5/10	66		38	31– 44	6	
P 5/6	30	6– 54	20	4– 35	40	39– 40
P 2/2	47	28– 65	270		67	
N8/ to 9/			32	22– 42		
N5/ to 6/			127	75–179		
N5/ to 4/			22	0– 43		
N2/ to 3/			11	0– 21		
Mean	44.8		60.2		30.4	

results. It shows percentage departures of the mean repeat limens from the original limens. The reliability results in general indicate that the determinations presented in the preceding tables are usually significant in the first decimal place and sometimes in the second.

Liminal Relations

It seemed desirable to analyze the data for possible systematic variations in the limens. The most obvious relationship to look for was that between magnitude of limen and magnitude of corresponding attribute. Figure 2 presents the difference limen for hue as dependent upon hue. The abscissa is discontinuous, representing the principal Munsell hues R, Y, G, B, and P. Each plotted point represents a mean limen for the several observers while a short horizontal line indicates the general mean of all limens for the given hue. The scatter of the liminal determinations for a given hue is due in large measure to the variety of values and chromas with which this hue was combined. Paying special attention to the locations of the general means, then, it is clear that the B limens are definitely high, the P limens are definitely low, and the remaining three vary insignificantly between. The high threshold for B confirms a similar finding for Munsell papers in 1939.[12]

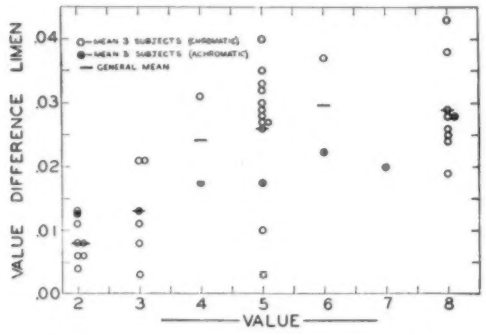


FIG. 3. Difference limen for value as a function of value.

Figure 3 shows the difference limen for value as related to value. The ordinate is in terms of the Munsell value unit while the abscissa is the value scale. Again each point is the mean of three thresholds and the short horizontals are the general means. This chart also shows marked scatter in the distribution at any value level. There is, in addition, some tendency for the size of the value limen to increase with increase in value.

Figure 4 exhibits the difference limen for chroma in relation to the dimension of chroma. The ordinate is scaled in the chroma unit while the abscissa represents the chroma scale. The distribution of plotted points at zero chroma represents absolute chroma thresholds. This chart depicts a very obvious tendency for the magnitude of the chroma limen to increase with increase in chroma.

Figure 5 portrays distribution curves for difference limens of hue, value, and chroma, respectively, and thus gives a general idea of the comparative variability and magnitude of the three. These curves include all the data with the exception of the reliability repeat measurements. In this chart the ordinate scale shows numbers of limens while the abscissa scale is that of the standard Munsell attributive unit. Thus, in case of hue limens, this scale is to be read in hue units; for chroma limens in chroma units; or for value limens in value units. The downward-pointing arrows indicate the general means of the hue, value, and chroma limens. The actual figures, together with the respective ranges of variation, are as follows: hue, 0.50 with range of 0.05–1.40; value, 0.021 with 0.002–0.049; chroma, 0.17 with 0.03–0.85.

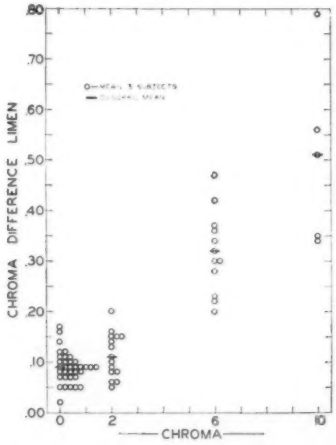


FIG. 4. Difference limen for chroma as a function of chroma.

DISCUSSION

Size of Scale Units

Some notion of the perceptual breadth of the Munsell scale steps can be secured from the general figures above. Thus, the attributive step or scale unit can be evaluated roughly in terms of the corresponding limen by merely dividing by the latter. Then the hue step is seen to be around 2 liminal differences, or just perceptible differences, broad; the chroma step averages about 6 just perceptible steps, broad; and the value step about 50 just perceptible steps. According to these figures, the attributive scales are far from perceptually equated at the liminal level, a condition long known to exist at superliminal levels.

The equational relation of the attributive scale units has been given for superliminal differences as: $1V=2C=3H$ (at $5C$).[10] The corresponding figures from our liminal data may be of some interest. Instead of using the general hue mean of 0.50 for this purpose, the hue mean at 6 chroma will correspond more closely to the hue at 5 chroma in the superliminal formula. This hue mean is 0.44 which yields a hue step approximately 2.3 just perceptible differences broad rather than 2. Reducing our just perceptible measures and substituting in the formula we have: $1V=8C=22H$ (at $6C$). The rank order of magnitude for the liminal and superliminal levels is seen to be the same, but the actual numerical amounts are very different.†

Value Scale Spacing

A. E. O. Munsell showed how the present data could be used to check the spacing of the neutral value scale by reapplication of the just noticeable difference method of calculating value steps as described in 1933.[5] Our figures, by this method, checked the seven intermediate steps of the value scale to within one or two-tenths of a unit, as shown below:

Value scale	2	3	4	5	6	7	8
Check	2.0	3.1	4.0	4.9	5.9	6.9	8.2

This check necessarily assumed agreement at levels 1 and 9. It seems remarkably close since planimetric estimates of area were involved and the liminal or j.n.d. methods employed were quite different. Of course this check merely verifies the spacing of the established scale without reference to the equality or inequality of the steps in that spacing.

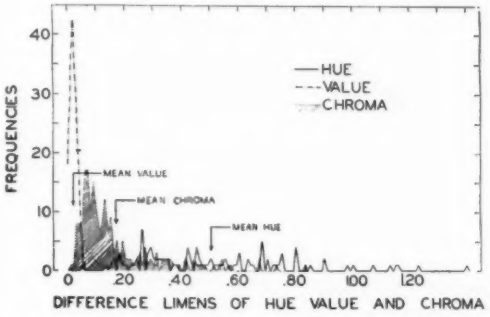


FIG. 5. Frequency distributions of difference limens of hue, value, and chroma.

Equality of Differences

The relatively large color differences perceived among the 400 standard samples of the Munsell system are known to approach the psychological ideal surprisingly well. In other words, these fairly obvious color differences which are supposed to look equal according to the 1929 notation, have in fact proved to approximate equality in visual appearance.

One cannot assume, however, that this desirable parallelism between size of scale differences and size of perceived differences holds for color differences of all magnitudes; in particular, one cannot assume that it would hold in the extreme case of the color differences which are perceived as minimal and which are represented by our color-difference limens. Thus, while liminal color differences must all be equal in the logical

† One of us has repeatedly observed that hue and chroma comparisons do not suffer as much as do lightness comparisons, when the samples compared are separated in space. In the present experiment all observations were made with juxtaposed samples, i.e., with larger and smaller disks on the same color wheel. This may account, in part at least, for the relative fineness of our lightness determinations, and the departure of our equation from that for superliminal color differences. Had we measured limens with samples separated by several inches, the lightness limens unquestionably would have been larger and our liminal equation for hue, value, and chroma units might have been considerably closer to the superliminal equation cited above.

sense that they are all just perceptible, we cannot be sure that they would be equal also as evaluated in terms of the Munsell scales. These scales might be correct for larger color differences but fail to be linearly summative with respect to smaller color differences.

Psychologists recognize at once that the problem is a form of the classical psychophysical problem of the validity of the difference limen as a scaling unit (reference 3, p. 143). We would like to use the difference limen as a scaling unit, though we can dispense with it if need be.

That the general indications are not too promising is immediately evident from the comparison of the equation for the attributive scales at the superliminal level with our equation at the liminal level. (Cf. second section preceding.) If the data were ideal and the difference limen were a suitable constant for a universal scaling unit, the two equations should be identical. Of course the very obvious difference must be due in part to the approximative nature of the equations and the data. Turning to analysis of the separate dimensions, we first note an encouragingly small range of variation in the magnitude of the value-difference limens (Fig. 5) and what is more important, no marked tendency to systematic variation in the value limens (Fig. 3). The hue limens appear to exhibit a much larger variability but this effect is due in considerable measure, to the psychologically finer hue scale. A systematic variation in the hue limens is apparent in Fig. 2, but this variation is also reflected in recent findings at superliminal levels, a fact which suggests that the remedy may lie in a correction of the hue scale. The chroma limens show an inescapable increase with increase in chroma (Fig. 4). Whether this Weber-like relationship can be reduced in some way seems doubtful.

But whether or not the difference threshold proves suitable for general use as a scaling unit, a color solid can be made exceedingly valuable by equi-spacing it at the utility level. The Munsell solid is already spaced at a useful level and the uniformity of this spacing is being improved by work now under way.[8] It seems probable, too, that the spacing may prove sufficiently equal over a considerable range of difference-sizes.

SUMMARY

Difference limens of hue, value, and chroma were obtained separately, under very favorable conditions of viewing, at various representative locations in the color solid. These data were secured by a form of the standard psychophysical method of constant stimulus differences; and the limens were computed as that amount of difference in hue, value, or chroma great enough to be perceived in 50 percent of many trials.

1. The general mean of the hue limens was 0.50 Munsell hue step with a range of 0.05–1.40; the mean of the value limens was 0.021 value step with a range of 0.002–0.049; the mean of the chroma limens was 0.17 chroma step with a range of 0.03–0.85. According to these figures, the Munsell hue step should average about 2 just perceptible differences in breadth, the chroma step about 6 just perceptible differences, and the value step about 50 just perceptible differences.

2. The graduations of the Munsell attributive scales are considerably farther from perceptual equivalence as evaluated with these liminal data than they are estimated to be at superliminal levels.

3. The hue limens appeared to be particularly high in the blue and particularly low in the purple. This effect is probably in part an artifact of a scale distortion in the blue to purple region.

4. The chroma limens exhibit a fairly regular increase in magnitude with increase in chroma. This result requires further investigation.

5. The value limens seemed to be in general relatively regular and remarkably small. These data were used to check the value scale by the method of Munsell, Sloan, and Godlove; and the results were in close agreement.

6. While logically the liminal differences must approximate perceptual equality at least as closely as the relatively large differences among the 400 standard samples, the former exhibit considerable variability as evaluated in terms of the Munsell scales. This indicates that the difference limen may not be an appropriate constant for use as a universal scaling unit.

7. During the course of the experimentation, the incidental discovery was made that the optimum period of adaptation for chroma discrimination is less than for value discrimination.

This suggests a fundamental difference in the two quantitative dimensions of color sensation.

ACKNOWLEDGMENTS

The writers are indebted to Louise Munsell and A. E. O. Munsell for invaluable assistance in collecting the data of this experiment, and to Mrs. A. H. Munsell for her interest and generous support.

BIBLIOGRAPHY

(1) R. G. Brown and L. B. Hoisington, "The dependence of hue on difference in tint," J. Gen. Psychol. **1**, 108 (1928).

(2) J. J. Glenn and J. T. Killian, "Trichromatic analysis of the *Munsell Book of Color*," J. Opt. Soc. Am. **30**, 609 (1940).

(3) J. P. Guilford, *Psychometric Methods* (McGraw-Hill, New York, 1936).

(4) Munsell Color Company, *Munsell Book of Color*, standard and abridged editions (Munsell, Baltimore, 1929).

(5) A. E. O. Munsell, L. L. Sloan, and I. H. Godlove, "Neutral value scales. I. Munsell neutral value scale," J, Opt. Soc. Am. **23**, 394 (1933).

(6) A. H. Munsell, *A Color Notation*, 9th Edition (Munsell Color Company, Baltimore, 1941).

(7) S. M. Newhall, "An interpolation procedure for calculating thresholds," Psych. Rev. **35**, 46 (1928).

(8) S. M. Newhall, "Preliminary report of the O.S.A. subcommittee on the spacing of the Munsell colors," J. Opt. Soc. Am. **30**, 617 (1940).

(9) Dorothy Nickerson, "Color tables for converting areas to terms of hue, brilliance, and chroma." Mimeographed report, Bur. Agri. Economics (U.S.D.A., 1935).

(10) Dorothy Nickerson, "The specification of color tolerances," Textile Research **6**, 505 (1936).

(11) D. J. Shaad and Harry Helson, "Group presentation in the method of constant stimuli as a time-saving device," Am. J. Psychol. **43**, 422 (1931).

(12) H. C. Smith, *Age Differences in Color Discrimination* (Johns Hopkins University dissertation, 1938).

(13) Symposium, "Color tolerance," Am. J. Psychol. **52**, 383 (1939).

Uniform Representation of the Characteristic Curves of Pure Silver Bromide Emulsions, Unsensitized and Sensitized†

LUDWIK SILBERSTEIN
Kodak Research Laboratories, Eastman Kodak Company, Rochester, New York
(Received June 8, 1942)

WHAT is usually called a *characteristic curve* or, in commemoration of Hurter and Driffield, *an H and D curve*, is the plot of the photographic density D, the common logarithm (Log) of the opacity produced by an exposure E, against Log E, the two variables, Log E and D, being treated as rectangular coordinates. The opacity or the reciprocal of transparency is a pure number and the exposure E is the radiant energy thrown upon the photographic plate or film per unit area and is, in recent work, measured and quoted in ergs/cm².

The general form of a characteristic curve will be familiar to most readers interested in photography. The gradient or slope of the curve,

$$g = dD/d \text{ Log } E$$

increases continuously (with increasing exposure) from zero up to a more or less distinct maximum which is commonly called "contrast" or "gamma," $\gamma = g_{max}$, and then decreases gradually to nil, while the curve ascends steadily, tending in normal conditions* asymptotically to a definite extreme or limiting density, usually denoted by D_m. This density is attained, of course, when all silver halide grains present are blackened. Theoretically, D_m corresponds to $E = \infty$ and thus to Log $E = \infty$. In most cases, the slope g of the curve varies in the neighborhood of g_{max} or γ very slowly, almost imperceptibly, so that a more or less extended stretch of the curve can scarcely be distinguished from a straight line, and this stretch is commonly designated by the practical photographer as the "straight-line portion" of the characteristic curve. (Within this stretch the density increases proportionally, or nearly so, to Log E.) Theoretically, however, assuming that not only D, but also its derivative g, are *continuous* functions of the exposure, the rectilinear stretch is infinitesimal, that is to say, the curvature vanishes only at a point, namely, where g attains its maximum, γ, and this point has to be conceived as the *inflection point* of the curve. This point, then, is characterized by $dg/dE = 0$. Its coordinates may be called the inflection-exposure and inflection-density and will, hereafter, be denoted by $\bar{E}$ and $\bar{D}$. In many cases, and notably for the unsensitized pure silver bromide emulsions, the inflection point is placed very nearly at half the height of the curve, $\bar{D} = \frac{1}{2}D_m$; the lower part of the curve, the so-called "toe," has the same shape as its upper part or "shoulder," and the inflection point is a true center of rotational symmetry. That is to say, a rigid rotation about this point through two right angles brings the whole curve into congruence with itself. In other cases, the characteristic curves show a pronounced asymmetry. In a certain class, however, of such cases, to be specified later, the nonsymmetrical curve can be split into two perfectly symmetrical ones. But even the symmetrical curves differ considerably in shape from each other in dependence on the peculiarities of the emulsions.

Since the early days of photographic theory, a variety of empirical mathematical formulae for the density as function of the exposure has been proposed and tried out. But, with a possible exception of C. C. Lienau and R. A. Houstoun's formula[1] and its modified form as retouched by J. H. Webb,[2] scarcely any of these formulae can claim to represent adequately the observed curves in their whole extent, from toe to shoulder. For the history of this subject the reader must be referred to F. E. Ross's monograph on *The Physics of the Developed Photographic Image*.[3]

The object of the present paper is to describe a comparatively simple sensitometric formula,

† Communication No. 859, Eastman Kodak Research Laboratories.

* Apart, that is, from the so-called "reversal," when the curve reaches a maximum height and descends again with increasing exposure.

[1] C. C. Lienau, J. Frank. Inst. **218**, 35–39 (1934); R. A. Houstoun, Phil. Mag. **21**, 1113 (1936).

[2] J. H. Webb, J. Opt. Soc. Am. **29**, 314 *et seq.* (1939).

[3] Monograph No. 5 of the Eastman Kodak Company (Van Nostrand, New York, 1924).

originally derived in 1939 by an asymptotic process from the quantum theory of exposure, but not hitherto published. This formula was since tried out extensively by the writer and turned out to represent very closely a great number of characteristic curves observed on various emulsions, under different development conditions. This fact of an extensive agreement with observations may justify perhaps the epithet "uniform" applied to the formula, and to its superpositions, in the title of the paper. At the same time, however, it must be said in advance that, in spite of its quantic origin, the close agreement of the formula with experiment cannot be considered as a decisive test in favor of the quantum theory, for reasons which will become clear hereafter. In fine, regardless of the manner of its derivation, the reader may take it, in the present connection at least, for an empirical or semi-empirical formula.

The reasoning which has led to the formula in question was as follows:

Let the exposure E amount to n quanta per unit area striking the plate haphazardly. Let a be the "size," i.e., the projected area, of a grain and ϵ its coefficient of photographically efficient absorption and, therefore $y=\epsilon an=cE$ (say), the mean or the probable number of efficient quanta per grain. Then the probability that the grain shall absorb m such quanta is, rigorously,

$$P_m(n)=p^m q^{n-m}\binom{n}{m}=p^m q^{n-m}\frac{n!}{m!(n-m)!},$$

where $p=\epsilon a$, $q=1-p$. The size a of the grain is supposed here to be a pure number, namely, the fraction of that area which has been taken above as "unit area," and since this area will be ultimately assumed to contain many grains, a and, *a fortiori*, $p=\epsilon a$ is a small fraction of unity, while n is a huge integer. The mean or the probable number of quanta absorbed by a grain is the sum of $mP_m(n)$ extended over all integer values of m and this is well known to be equal to np or y, as stated above.

If not only n, but also the numbers m and $n-m$ are great enough to warrant the application of Stirling's expression for the factorials, $n!=(2\pi n)^{\frac{1}{2}}\cdot n^n e^{-n}$, the rigorous probability leads, approximately, to

$$\frac{1}{(2\pi y)^{\frac{1}{2}}}\int_\alpha^\beta \exp\left[-\frac{(m-y)^2}{2y}\right]dm$$

as the expression for the probability that a grain shall absorb α to β quanta. This expression will be familiar to most readers in connection with the usual statement of the well-known theorem of Bernoulli.[4]

Suppose, now, that a grain becomes developable whenever it absorbs at least r quanta (i.e., r or more). Then the probability that the grain shall be made developable by an exposure $y=a\epsilon n=cE$ will be

$$P(y,r)=\frac{1}{(2\pi y)^{\frac{1}{2}}}\int_r^\infty \exp\left[-\frac{(m-y)^2}{2y}\right]dm$$

or, in terms of the error-function,

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}du,$$

$$P(y,r)=\frac{1}{2}\left\{\Phi(\infty)-\Phi\left(\frac{r-y}{(2y)^{\frac{1}{2}}}\right)\right\}$$

and since Φ is an odd function of its argument and $\Phi(\infty)=1$,

$$P(y,r)=\frac{1}{2}\left\{1+\Phi\left(\frac{y-r}{(2y)^{\frac{1}{2}}}\right)\right\}.$$

If there are a great many equal grains spread over our unit area, $P(y,r)$ will give also the fraction of all these grains made developable by the exposure y. And since the photographic density D is proportional to the number of grains made developable and, in particular, D_m proportional to the total number of grains over unit area,[5] the resulting sensitometric formula is

$$D=D_m\cdot P(y,r)=\frac{D_m}{2}\left\{1+\Phi\left(\frac{y-r}{(2y)^{\frac{1}{2}}}\right)\right\}. \quad (1)$$

It is this formula which will occupy our attention in the present paper.

It replaces the more accurate quantum

[4] Its simplified derivation is given in a note published in Phil. Mag. **26**, 223 (1938). In the present case, $np=y$ and $q=1-p\doteq 1$ in the radical.

[5] Cf. J. Opt. Soc. Am. **28**, 443 (1938).

formula

$$\frac{D}{D_m}=1-e^{-y}\left\{1+y+\frac{y^2}{2!}+\cdots+\frac{y^{r-1}}{(r-1)!}\right\} \quad \text{(A)}$$

used in my previous publications.

Rigorously, the Stirling expression for the factorial $r!$ and thus also the formula (1) would be valid asymptotically, for $r\to\infty$, and may be appropriately referred to as the *asymptotic* quantum formula. Practically, however, the Stirling expression is a fair approximation to $r!$, and thus also (1) to (A), when r amounts only to several units.[6] To what extent the asymptotic formula (1) approximates the correct quantic densities even for moderate values of r may best be seen from the following set of densities for $y=r$ (inflection point). By (A) we have, to three decimals,

r =	1	2	3	4	5	6	7
$D(r)/D_m$ =	0.632	0.594	0.577	0.566	0.559	0.554	0.550,

while, by (1), $D(r)/D_m=\frac{1}{2}$. For $r=6$ or 7, the difference amounts only to 10 percent, at the inflection point, and decreases gradually down to zero at higher or lower exposures. For values of r, however, as low as 2 or 1 the formula (1) deviates from (A) considerably. In fine, for such low values of the number r the asymptotic formula (1) can no longer be considered as an approximate implication of the quantum theory of exposure, and still less so for fractional values, $r<1$, when the formula (A) becomes meaningless.

Now, while none of the (A) formulae, as they stand, is adequate to cover satisfactorily any of the observed characteristic curves in their whole extent, the asymptotic formula (1) has turned out to represent very closely indeed a great variety of such curves, but always with surprisingly *low values* of the parameter r derivable directly from, and determined uniquely by, the main characteristics of the observed curve (D_m and γ; *vide infra*). In no case hitherto investigated, in fact, did r exceed 2.7 or 2.8, and in the great majority of cases was much less than that, descending to fractional values as low as 0.30 or 0.25. This, then, is the reason why the close agreement of the formula (1) with experimental measurements cannot be invoked as a proof of the quantum theory of exposure.[7] The formula, moreover, has been derived above under the explicit assumption that all grains are equal, while it has actually turned out to fit the curves of emulsions with a wide range of grain sizes as closely as the curves of almost uniform emulsions. The number r called for by a given emulsion would thus have to be viewed as a kind of average of the individual characteristics of the grains composing it.

We shall, accordingly, adopt here a purely phenomenological attitude and, forgetting the original meaning of r (as number of quanta required for developability) consider it merely as a *parameter* of our formula to be determined from some directly measurable properties of the observed characteristic curve.

Before passing on to the applications of the proposed formula, it will be well to examine its main properties.

GENERAL PROPERTIES OF THE ASYMPTOTIC FORMULA

The formula

$$\frac{D}{D_m}=\frac{1}{2}\left\{1+\Phi\left(\frac{y-r}{(2y)^{\frac{1}{2}}}\right)\right\} \quad (1)$$

contains in all three parameters: the coefficient of proportionality of our independent variable to the exposure, $c=y:E$, the extreme density D_m, and the number r. The parameter c affects only the position of the corresponding curve; its change produces only a rigid shift of the whole curve along the Log E axis. The parameter D_m is a common factor of all ordinates, so that the shape of the curve is influenced by the parameter r only.

Needless to say, for $y=0$, $\Phi=-1$, $D=0$ and for $y\to\infty$, $\Phi=1$ and $D=D_m$.

The most characteristic point of the curve is its midpoint,

$$y=r, \qquad D(r)=\tfrac{1}{2}D_m.$$

[6] E.g., for $r=12$ the difference is about $0.007r!$ and even for $r=6$ amounts to $0.014r!$ only.

[7] The situation is, perhaps, modified by the frequent occurrence of "clumps" of grains which are well known to have the property that one only of the grains need be effectively exposed in order that all grains forming the clump shall be made developable. This would imply an apparent lowering of the r value conceived as an average. At best, the question at issue could for the present be qualified as an open question whose clinching, however, is not one of the purposes of this paper.

We know beforehand that this will also be its *inflection point*, for such was $y=r$ for all the quantic curves (A) corresponding to $r=1, 2, 3,$ etc., of which (1) is but a limit. This will be verified presently as a differential property of (1) itself.

The midpoint, $y=r$, has also another remarkable property. In fact, consider any two points *equidistant* from it along the Log y axis or

$$y=mr \quad \text{and} \quad y^*=\frac{1}{m}r,$$

where m is any positive number whatever. Let u be the argument of the error function in (1). Then, for the first point,

$$u=\frac{(m-1)r}{(2mr)^{\frac{1}{2}}}=(m-1)\left(\frac{r}{2m}\right)^{\frac{1}{2}}$$

and for the second point,

$$u^*=\frac{(1/m-1)r}{(2r/m)^{\frac{1}{2}}}=(1-m)\left(\frac{r}{2m}\right)^{\frac{1}{2}}=-u.$$

Hence,

$$D(y)=\frac{D_m}{2}\{1+\Phi(u)\} \quad \text{and} \quad D(y^*)=\frac{D_m}{2}\{1-\Phi(u)\}$$

or

$$D(y)+D(y^*)=D_m.$$

This holds for every pair of points y and $y^*=r^2/y$.

Thus the curve is rotationally *symmetrical* and the point $y=r$ is the center of its symmetry. A rigid rotation of the curve about this point through two right angles brings the whole curve into congruence with itself, its shoulder and toe being interchanged.[8] In view of this property, this point will henceforth be called briefly *the center* of the curve and will be denoted by C.

Let us now consider the differential properties of the curve (1).

Writing for brevity $u=(y-r)/(2y)^{\frac{1}{2}}$, recalling the properties of the error-function, and denoting derivatives with respect to y by dashes, we have

$$\frac{D'(y)}{D_m}=\frac{u'}{\sqrt{\pi}}\exp(-u^2).$$

The gradient of the curve is, by definition, $g=yD'(y)/M$, where $M=\text{Log } e$, so that

$$\frac{M}{D_m}g(y)=\frac{1}{\sqrt{\pi}}yu'\exp(-u^2). \tag{2}$$

The inflection point $\bar{y}$ is determined by $g'(y)=0$ or, after removal of the common factor $\exp(-u^2)$ (whose vanishing gives the obvious roots $y=0$ and $y=\infty$), $u'+y[u''-2uu'^2]=0$, and since

$$u'=\frac{y+r}{2(2y^3)^{\frac{1}{2}}}, \quad u''=-\frac{y+3r}{4(2y^5)^{\frac{1}{2}}},$$

this equation is reduced to

$$(y-r)[(y+r)^2-y]=0,$$

a cubic for $\bar{y}$. One root is

$$\bar{y}=r, \tag{3}$$

as announced above, while the remaining two roots are, for all values of r greater than $\frac{1}{4}$, complex conjugate.[9]

Thus (unless $r<0.25$, which is seldom the case) the curve has a *unique* inflection point, $\bar{y}=r$.

At the same time, Eq. (2) gives for γ or $g(r)$

$$M\gamma/D_m=(r/2\pi)^{\frac{1}{2}},$$

whence

$$r=2\pi\left(\frac{M\gamma}{D_m}\right)^2, \tag{4}$$

a simple expression for the parameter r in terms of the directly observable attributes γ, D_m of the curve which is to be represented by the formula (1).

In view of the theorem proved a moment ago, only *symmetrical* curves (or nearly so) can be covered by that formula, which will henceforth be written briefly

$$D=D_m\cdot P(y, r). \tag{1a}$$

A table of the function

$$P(y, r)=\frac{1}{2}\left\{1+\Phi\left(\frac{y-r}{(2y)^{\frac{1}{2}}}\right)\right\}$$

[8] We may mention that none of the curves (A) is strictly symmetrical, but they tend to become so with ever increasing r.

[9] If $r<\frac{1}{4}$, these roots, y_1, $y_2=\frac{1}{2}[1-2r\pm(1-4r)^{\frac{1}{2}}]$, are not only real, but both positive and give actual *maxima*, $g(y_1)=g(y_2)$, of the slope curve. Since $y_1y_2=r^2$, these two maxima are placed symmetrically with respect to the central ordinate $g(r)$ which now is a *minimum* (although, even for $r=\frac{1}{4}$, only 10 percent lower than either maximum). When $r=\frac{1}{4}$, the two maxima coalesce with $g(r)=\gamma$ which then is a "flat" maximum and which becomes more and more "sharp" with increasing r.

which will be found sufficient for all actual cases is given at the end of the paper.

The process of adapting the formula to an observed curve is very simple and rapid. The values of γ and D_m being read off the given curve (the latter, D_m, if need be, after a slight extrapolation of the highest measured densities), the value of the parameter r is readily determined by (4). The position of the curve (1a) is best fixed by placing its center $C(y=r)$ upon the observed curve, that is to say, simply at the observed height $D=\frac{1}{2}D_m$. This gives the precise inflection exposure $\bar{E}$, say in ergs/cm² (which otherwise could be read off but vaguely), and herewith also the last parameter c appearing in $y=cE$. In fact, $c=r/\bar{E}$. By the way in which r has been determined the two curves have now a common inflection tangent at C and thus a contact of the third order. It suffices then to calculate D by (1a) for a few higher steps, $D(2r)$, $D(4r)$, etc., the fewer the greater r. The densities below the center are then codetermined by the said symmetry, *viz.*, $D(r/2)=D_m-D(2r)$, $D(r/4)=D_m-D(4r)$, and so on. As will be seen from the P table, the extreme density is practically reached in three such steps above the center for $r=1$, and in five steps for $r=0.30$ which is about the lowest r-value called for by any emulsion. And as many steps below C lead, of course, almost to $D=0$. For $r=3$, which is a little above the highest parameter ever encountered in practice, $D(4r)$ is only $\frac{1}{2}$ percent below D_m and, correspondingly, $D(r/4)=0.005D_m$, which in view of the p.e. of the density measurements may be practically confounded with zero.

The symmetry of a curve, with Log E as abscissa, is most directly expressed by using, instead of E or $y=cE$, as independent variable,

$$x=\text{Log } E-\text{Log } \bar{E}=\text{Log } (E/\bar{E})=\text{Log } (y/\bar{y}).$$

In terms of x, any symmetrical curve whatever is obviously expressed by

$$D=\frac{D_m}{2}\{1+f(x)\}, \qquad (5)$$

where $f(x)$ is some *odd* function of x, and thus $f(0)=0$, and $\gamma=\frac{1}{2}D_m f'(0)$, while $f(\infty)=1$.

To reduce our formula (1) to this form, notice that the argument of the error-function Φ can be written

$$u=(r/2)^{\frac{1}{2}}[(cE/r)^{\frac{1}{2}}-(r/cE)^{\frac{1}{2}}]$$
$$=(r/2)^{\frac{1}{2}}[(E/\bar{E})^{\frac{1}{2}}-(\bar{E}/E)^{\frac{1}{2}}]$$
$$=(r/2)^{\frac{1}{2}}[10^{x/2}-10^{-x/2}],$$

or

$$u=(2r)^{\frac{1}{2}}\sinh\frac{x}{2M}.$$

Thus the formula becomes

$$D=\frac{1}{2}D_m\left[1+\Phi\left((2r)^{\frac{1}{2}}\sinh\frac{x}{2M}\right)\right], \qquad (1b)$$

where r has still the value (4). Needless to say, Φ has all the properties of $f(x)$ stated above. The gradient is now, simply, $g(x)=D'(x)$, and $g'(x)$ vanishes at $x=0$, the inflection point. Differentiating once more, one readily finds that the sign of $g''(0)$ is that of $\frac{1}{4}-r$. Hence, if $\gamma=g(0)$ is to be a true maximum of the gradient, r cannot be less than $\frac{1}{4}$. If $r<\frac{1}{4}$, then $g(0)$ is actually a minimum, bordered symmetrically by two maxima. (See footnote 9.)

SUPERPOSITION OF TWO TERMS OF THE FORM (1). ASYMMETRICAL CURVES

As has already been mentioned, the characteristic curves of unsensitized, pure silver bromide emulsions are all (rotationally) symmetrical and can be very closely represented by the formula (1) or (1a), identical with (1b). When, however, any of these emulsions is *sensitized* by some special process or other, the symmetry of its curve is destroyed, and the curve can, therefore, no longer be represented by a single term $D_mP(y, r)$.

Now, as has been remarked by Mr. A. P. H. Trivelli of these Laboratories (to whom I am obliged for all the experimental data used in this paper), there are good reasons to suppose that the grains of an emulsion are differently affected by the sensitizing process and that in this respect, the whole emulsion might, perhaps, be considered summarily as consisting of two groups of grains or of two components, each behaving sensitometrically as the original emulsion, but endowed with different parameter values. If such were the case, the asymmetrical curves of the sensitized emulsion ought to be expressible by *two terms* of the form (1a), that is to say, by

$$D=A_1P(y_1, r_1)+A_2P(y_2, r_2),$$

where $A_1+A_2=D_m$, and $y_i=c_iE$. Such, in fact, has turned out to be the case for all samples so far investigated. After some trials, moreover, it was found that the best superposition of terms is what may be called a *two-to-one* mixture ($A_1=2A_2$), so that the working formula became

$$D=D_m\{\tfrac{2}{3}P(y_1, r_1)+\tfrac{1}{3}P(y_2, r_2)\}. \qquad (5)$$

It remained to determine the values of the four parameters r_1, r_2 and c_1, c_2 (implied in y_1, y_2) or, equivalently, the position, Log $\bar{E}_1$, Log $\bar{E}_2$, of the centers of symmetry C_1 and C_2 of the two component curves, so as to fit the observed asymmetrical curve. The limiting densities, $A_1=\frac{2}{3}\cdot D_m$ and $A_2=\frac{1}{3}\cdot D_m$, of the component curves being known, the parameters r_1, r_2 are, in virtue of (4), reducible to the gammas of these curves,

$$r_1=2\pi\left(\frac{3}{2}\frac{M\gamma_1}{D_m}\right)^2, \quad r_2=2\pi\left(\frac{3M\gamma_2}{D_m}\right)^2, \qquad (4a)$$

but γ_1, γ_2 themselves are unknown and their relation to the resultant γ derivable from (5) is of a very intricate nature. Theoretically, the four parameters, c_1, c_2 and r_1, r_2 (or γ_1, γ_2), would be determined by four points of the observed curve. But the solving of the corresponding system of four equations containing, through P, the transcendental function Φ of the unknown arguments, would be an extremely laborious task.

To obviate these difficulties, the following method has been adopted.

A preliminary investigation of several cases having shown that the toe of the observed curve can be closely enough represented by the single term

$$D_1=\tfrac{2}{3}\cdot D_mP(y_1, r_1), \qquad (5.1)$$

the contrast (γ_1) of this partial curve has been determined by identifying it with the gradient of the observed curve at a point well below its midpoint ($\frac{1}{2}D_m$). In practice, one or two trials turned out to lead to the best choice of γ_1. From this value of γ_1 the parameter r_1 is directly computed by the first of (4a). A few points, $y_1=r_1$, $\frac{1}{2}r_1$, $\frac{1}{4}r_1$, etc., of the curve (5.1) are then determined with the aid of the P table and one of these points is placed on the observed curve, this point being so chosen that all the lower points also fall upon or very closely to the observed curve. (In most cases it turned out to be the point $y_1=\frac{1}{2}r_1$ and sometimes the center C_1 itself, $y_1=r_1$.) Thus also the position of the D_1 curve is fixed. Let it be drawn completely, with the aid of the P table. By construction, the lower part of this curve will follow closely the toe of the observed curve, say up to $y_1=\frac{1}{2}r_1$ or r_1, and beyond that will fall below the observed curve, deviating from it gradually more and more, the limit of the deviation being $\frac{1}{3}D_m$. We now mark, for a number of points, the difference $D_{obs}-D_1$, as ordinate of these points. From the smooth curve

$$D_{obs}-D_1, \text{ Log } E \qquad (B)$$

drawn through these points we read off its gamma and, in order to satisfy the requirement $D_{obs}-D_1=D_2$, we take this gamma for γ_2, which gives us the parameter $r_2=2\pi(3M\gamma_2/D_m)^2$. The curve

$$D_2=\tfrac{1}{3}\cdot D_mP(y_2, r_2) \qquad (5.2)$$

can now be computed, and its center C_2 (with density $\frac{1}{6}\cdot D_m$) being placed on the curve (B), the whole curve (5.2) can be drawn on the same chart in its proper position. This completes the process which is by no means laborious.

If the points of the curve (5.2) lie close to the curve (B), the superposition of the two component curves, D_1+D_2, will lie as closely to the observed curve. For the differences $D_2-(D_{obs}-D_1)$ are manifestly identical with the deviations

$$\Delta=(D_1+D_2)-D_{obs},$$

which ultimately interest us.

Some details of this general description of the computation process will become more lucid from the concrete examples to which we now turn.

I. SYMMETRICAL CURVES OF UNSENSITIZED EMULSIONS

Since the process of adapting the single-term formula (1) to a symmetrical observed curve is straightforward and simple enough, we shall describe it in some detail only for the first of our concrete examples. In the case of all further examples it will suffice to give the final results only and the corresponding graph or figure. The abscissa of each figure is Log E, with the exposure E, throughout at constant intensity and varying exposure time, expressed in ergs/cm^2, unless otherwise stated in the diagrams. The

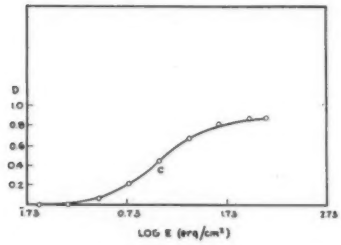


FIG. 1. N4X-7229, unfinished. 2′ development in ferrous oxalate.

development time, in minutes, will be briefly denoted by t.

1. Emulsion N4X-7229 ("unfinished"), $t=2$ in Ferrous Oxalate

The observed curve (Fig. 1) shows, nearly enough, the limiting density $D_m=0.87$ and $\gamma=0.80$. Whence, by formula (4), to two decimals, $r=1.00$. Thus formula (1) or (1a) becomes

$$D=0.87P(y, 1.00).$$

This, with the aid of the P table, gives the following set of densities:

$y/r=$	1	2	4	8	16	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
$D=$	0.43_5	0.66	0.81	0.86	0.87	0.21	0.06	0.01	0.00

The corresponding points are marked in Fig. 1 by eyelets. The center $C(y=1.00)$ being placed on the observed curve, which gives Log $\bar{E}=1.05$, the remaining points are all very close to the observed curve. The greatest deviation $\Delta=D-D_{\text{obs}}$ amounts barely to -0.02 low down in the toe and to $\Delta=0.02$ two steps above the center. These small deviations scarcely exceed the p.e. of the density measurements.

Equally close representations by $D=D_mP(y, r)$ have been found for the same emulsion with $t=3$, 4, 6, 9 and 18 min. in ferrous oxalate, nay, with much the same values of the parameters r and Log $\bar{E}$, the former oscillating between 0.88 and 1.00 and the latter between 1.03 and 1.07. Of these five cases, it may suffice to quote one only.

2. Emulsion N4X-7229, $t=9$ in Ferrous Oxalate

The observed curve (Fig. 2) is represented throughout very closely by $D=1.02P(y, 0.98_5)$, with center C placed again at Log $\bar{E}=1.05$. The greatest deviation is $\Delta=-0.01_5$ only.

3, 4. Emulsion S-124-1, $t=5$ and 8 in DK-50

The formulae, are, respectively,

$$D=2.16P(y, 0.325) \quad \text{and} \quad D=2.64P(y, 0.305).$$

See Fig. 3, which shows a perfect agreement of the calculated points with both experimental curves.

5, 6. Emulsion S-124-1, $t=5$ and 8 in DK-50

These two curves are represented, respectively, by

$$D=1.76P(y, 0.31) \quad \text{and} \quad D=2.40P(y, 0.28)$$

very closely indeed, as will be seen from Fig. 4.

7, 8. Emulsion N4X-3131, $t=6$ and 8 in D-19

The solid curve drawn in Fig. 5 corresponds to

$$D=1.60P(y, 1.10)$$

and represents closely enough both sets of measured densities, marked by crosses for $t=6$ and by eyelets for $t=8$.

9

Some eighty more experimental characteristic curves have been similarly investigated in comparison with theoretical ones of the form (1). In order to save space in the columns of this Journal, the results found may be represented summarily (and not less instructively, perhaps) in a single diagram, namely, with D/D_m as ordinate and the error function Φ itself (a known function of the exposure, containing c, r as parameters) as abscissa. All the particular theoretical curves are then condensed in a unique straight line,

$$D/D_m=\tfrac{1}{2}(1+\Phi), \qquad (6)$$

extending horizontally from $\Phi=-1$ up to

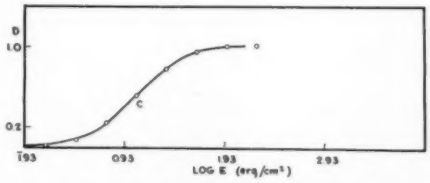


FIG. 2. N4X-7229, unfinished. 9′ development in ferrous oxalate.

$\Phi=1(E=0$ to $\infty)$ and vertically from $D/D_m=0$ up to 1, and condensing, as it were, in its mid-point the inflection points of all those curves. The observed points (with their coordinates derived from the original curvilinear diagrams) are represented by dots which fall more or less closely to this straight line. Such a diagram, summarizing a good number of results found for a variety of emulsions (*viz.*, an x-ray emulsion, Cine Positive, and some ten more emulsions, at different development conditions), is given in Fig. 6 which ought to be self-explanatory. It contains nearly two hundred observed "points" or dots, not selected ones, of course. These points cluster throughout very closely about the straight line (6). It will be seen that even at the upper part of the shoulder, $\Phi=0.85$ to 1.00, considerable deviations are rather sporadic only. This diagram seems to give a good cumulative evidence in favor of the formula $D=D_mP(y, r)$ as a faithful representative of the symmetrical variety of the observed curves.

II. ASYMMETRICAL CURVES OF SENSITIZED EMULSIONS

We shall now consider a few applications of the two-term formula (5),

$$D=\tfrac{2}{3}D_mP(y_1, r_1)+\tfrac{1}{3}D_mP(y_2, r_2)\equiv D_1+D_2.$$

In order not only to show its applicability to non-symmetrical observed curves, but also to bring out the effect of the process of sensitizing, it will be best to consider first an emulsion whose unsensitized sample has already been investigated in Section I.

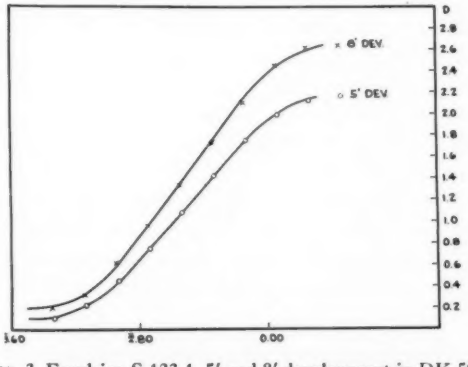


FIG. 3. Emulsion S-123-1. 5′ and 8′ development in DK-50.

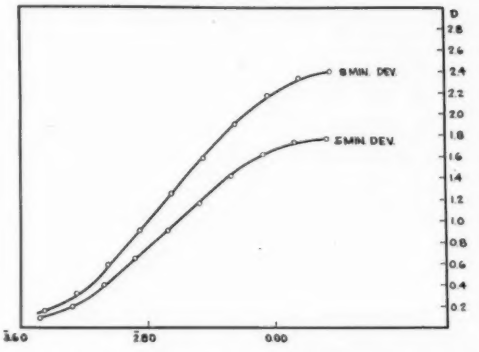


FIG. 4. Emulsion S-124-1. 5′ and 8′ development in DK-50.

10. Emulsion N4X-7229 Sensitized, $t=2$ in Ferrous Oxalate

This is the same emulsion as that treated in the examples 1 and 2, above, but sensitized to form silver sulfide specks.

Its observed curve is the upper solid curve drawn in Fig. 7. Its pronounced asymmetry can be seen at a glance. The limiting density, actually reached and maintained for a good stretch of the curve, is $D_m=1.30$. This is split, in near accordance with formula (5), into $0.87+0.43$, so that the two component curves become

$$D_1=0.87P(y_1, r_1) \qquad \text{and} \qquad D_2=0.43P(y_2, r_2).$$

Having adopted for γ_1, after a little trial, the gradient at the lower part of the observed curve, $(0.32-0.17):0.30$, i.e., $\gamma_1=0.50$, we find, by (4a), $r_1=0.39$ and thus $D_1=0.87P(y_1, 0.39)$, which gives the following set of densities D_1:

y_1/r_1	= 1	2	4	8	16	32	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
D_1	=0.43$_5$	0.58	0.72	0.81	0.86	0.87	0.29	0.15	0.06	.01	.00
$D_{obs}-D_1$	=0.05	0.16	0.28	0.36	0.40	0.43	0.00	.00	.00	.00	.00

The D_1 points are marked by dots in Fig. 7. One of these, $D_1=0.29$, being placed on the observed curve, the dots below it fall all perceptibly on this curve, while the center C_1 and the successive dots ($y_1/r_1=1$, 2, 4, etc.) fall gradually away from the observed curve, as shown by the values of $D_{obs}-D_1$, given above. The placing of the D_1 dots on the chart (or the correlation of the y_1 and E scales) fixes, at the same time, the position of the center C_1, marked by an eyelet, giving for its abscissa,

$$\text{Log } E_1=0.47.$$

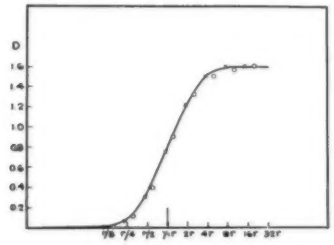


FIG. 5. N4X-3131. Developed in D-19.

The differences $D_{obs}-D_1$ are now plotted, *in situ*, against Log E, which gives the curve (so superscribed) in the lower part of the chart. This curve, which practically coincides with the axis up to about Log $E=0.17$, is, unlike the resultant observed curve, strikingly symmetrical and turns out to be closely expressible by a single term of the form (1). From its limiting density and its gamma one readily finds $r_2=0.925$, always by (4a). This gives, for the required second component,

$$D_2=0.43P(y_2, 0.925).$$

The points computed by this formula are marked by eyelets. The central point $C_2(D_2=0.21_5)$ being placed on the $(D_{obs}-D_1)$ curve, which gives for its position

$$\text{Log } \bar{E}_2=0.89,$$

the remaining eyelets or D_2 points array themselves very closely along the $(D_{obs}-D_1)$ curve. Manifestly, the deviations of the eyelets from this curve are identical with the deviations of the corresponding values of the sum D_1+D_2, marked by crosses, from the total observed densities. The greatest of these deviations, $\Delta=D_1+D_2-D_{obs}$, amounts barely to 0.02. In fine, the observed asymmetrical curve is very closely represented in its whole extent by

$$D=0.87P(y_1, 0.39)+0.43P(y_2, 0.925),$$

the sum of two symmetrical curves. The coefficients c_1, $c_2=r_1/\bar{E}_1$, $r_2/\bar{E}_2$ implied in y_1, y_2 are codetermined by the position of the centers C_1, C_2 of these curves,

$$\text{Log } \bar{E}_1=0.47, \text{ Log } \bar{E}_2=0.89.$$

For the unsensitized emulsion, with exactly the same development, we have found (Ex. 1) for

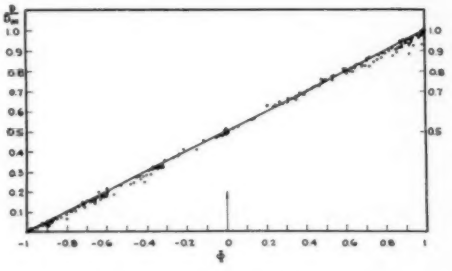


FIG. 6. Observed D/D_m plotted against Φ.

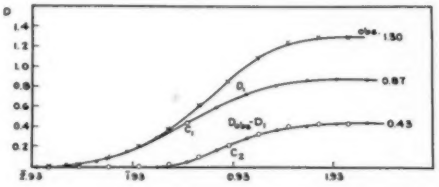


FIG. 7. N4X-7229, sensitized. 2′ development in ferrous oxalate.

the center of the single symmetrical curve

$$\text{Log } \bar{E}=1.05.$$

The reciprocal of the inflection exposure, at which half the limiting density is reached, may be considered as a measure of the *sensitivity* of the emulsion, under given development conditions, $s=1/\bar{E}$. Similarly, for the two components of the sensitized emulsion, s_1, $s_2=1/\bar{E}_1$, $1/\bar{E}_2$. Thus we have, in cm²/erg, for the unsensitized emulsion, $s=0.089$, and for the primary[10] and the secondary components of the sensitized emulsion,

$$s_1=0.34, \qquad s_2=0.13$$

or $s_1=3.8s$ and $s_2=1.45s$. The process of sensitizing produces thus a kind of bipartition of the emulsion, increasing the sensitivity of its main part about four times and that of the remainder only 1.5 times. The grains of the original emulsion may well be considered as endowed with different susceptibilities to the sensitizer dependent on some difference in their individual peculiarities (such as size) which is revealed only by the sensitizing process. Whether the greater susceptibility belongs, on an average, to the larger grains or not, has not, so far, been ascertained.

An equally close representation by the two-

[10] We may call it so since it makes up two-thirds of the whole emulsion with regard to limiting density.

term formula has been obtained for the curves of the same emulsion for five more development times, $t=3$ to 18 min. in ferrous oxalate, as well as for 3 to 12 min. in DK-50. Of all these cases, it will suffice to quote here only the following one, since it is directly comparable with that treated in Ex. 2.

11. Same Emulsion, but with $t=9$ in Ferrous Oxalate

Proceeding exactly as in 10, one finds the superposition of two symmetrical curves

$$D=1.06P(y_1, 0.345)+0.53P(y_2, 1.69),$$

which agrees with the observed curve as closely as in the last case. All details can be seen from Fig. 8, where all symbols and marks have the same meaning as in Fig. 7.

The two centers of the symmetrical component curves are now placed at

$$\text{Log } \bar{E}_1=0.42 \qquad \text{and} \qquad \text{Log } \bar{E}_2=0.79_5,$$

while the unique center of the unsensitized emulsion, with exactly the same development (Ex. 2) had the position

$$\text{Log } \bar{E}=1.05.$$

Thus the sensitivities are

$$s=0.089; \qquad s_1=0.38, \qquad s_2=0.16; \quad (t=9').$$

s is the same as, and s_1, s_2 are but slightly greater than for $t=2$ min. The "bipartition" or the selective effect of the sensitizer is thus, nearly enough, independent of the development time.

12. Emulsion N4X-7233, $t=2$ and 3 in DK-50

This is the same emulsion as in Ex. 1, but sensitized with *active gelatin*. Its two curves (Figs. 9 and 10 for $t=2$ and 3, respectively) are distinctly asymmetrical. The first is closely represented by

$$D=1.05P(y_1, 1.55)+0.53P(y_2, 1.55)$$

with centers placed at

$$\text{Log } \bar{E}_1=0.29 \qquad \text{and} \qquad \text{Log } \bar{E}_2=0.97$$

and $\Delta_m=0.03_5$ as greatest deviation, and the second is equally well represented by

$$D=1.19P(y_1, 1.64)+0.59P(y_2, 1.64)$$

with

$$\text{Log } \bar{E}_1=0.25 \qquad \text{and} \qquad \text{Log } \bar{E}_2=0.91$$

and $\Delta_m=\pm 0.04$. An interesting peculiarity of this case is that, to our approximation, $r_1=r_2$ for either development time, so that the two component curves are essentially the same, differing only as to vertical scale and position. The sensitivities of the components are

$$s_1=0.51, \quad s_2=0.11 \quad \text{for} \quad t=2 \text{ min. in DK-50}$$

and

$$s_1=0.56, \quad s_2=0.12 \quad \text{for} \quad t=3 \text{ min.}$$

In fine, the sensitivity of the primary component is about *five* times that of the secondary com-

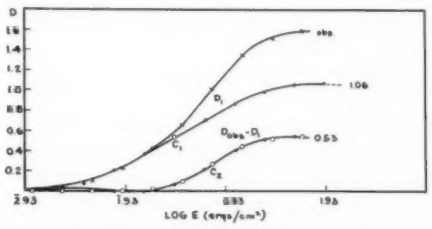


FIG. 8. N4X-7229, sensitized. 9′ development in ferrous oxalate.

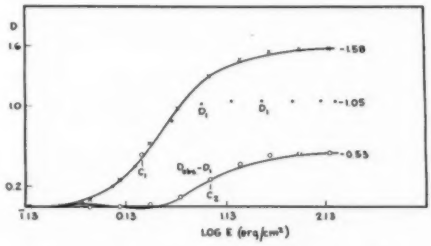


FIG. 9. N4X-7233, finished with active gelatin. 2′ development in DK-50.

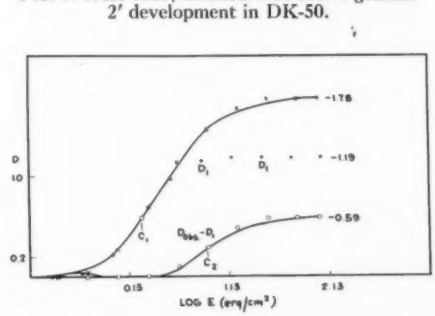


FIG. 10. N4X-7233, finished with active gelatin. 3′ development in DK-50.

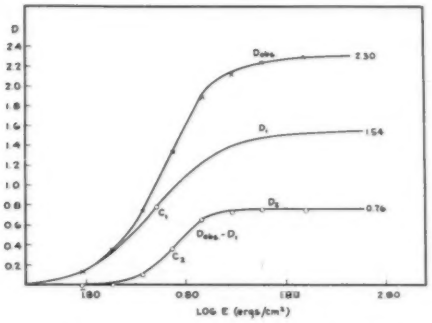


FIG. 11. Emulsion I. XX D_1+D_2. $D_2=0.76P(0.449E, 2.05)$. $D_1=1.54P(0.158E, 0.50)$.

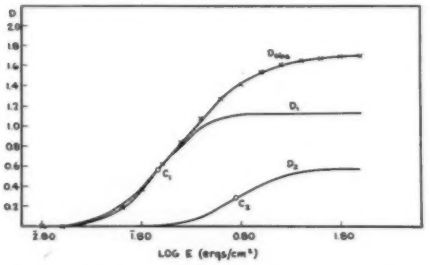


FIG. 12. XXX D_1+D_2. C_1, C_2, centers (inflection points). Emulsion II. $D_1=1.13P(0.975E, 0.93)$. $D_2=0.57P(0.185E, 1.04)$.

ponent and the effect of the prolonged development on s_1, s_2 is almost negligible (although the total extreme densities for $t=2$ and 3 differ from each other as much as 1.58 from 1.78). The symmetrical curves of the unsensitized emulsion developed in DK-50 (not reproduced here) show the inflection exposure Log $\bar{E}=1.02$ for the shorter, and 1.00 for the longer development and thus the sensitivity $s=0.095$ and 0.10, not differing much from that found with ferrous oxalate as developer.

13. A Set of Seven Converted (Ammoniacal) Emulsions

These emulsions differ only as to their precipitation time (t_p). Their investigation, through the two-term formula, has afforded the additional interest of ascertaining how the parameters of that formula, and especially the two sensitivities s_1, s_2 depend on the average grain size ($\bar{a}$) of an emulsion which increases steadily with t_p. The precipitation time affects not only $\bar{a}$, but also the standard deviation σ and thus the shape

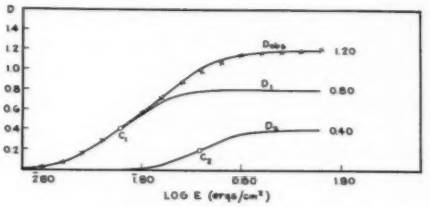


FIG. 13. XXX D_1+D_2. Emulsion III.

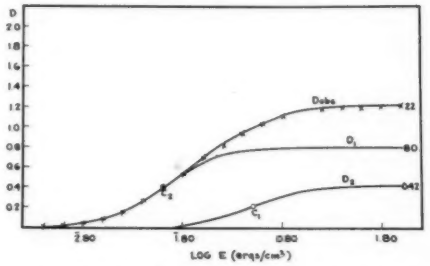


FIG. 14. XXX D_1+D_2. Emulsion IV.

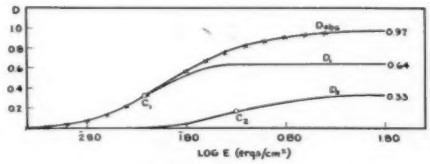


FIG. 15. XXX D_1+D_2. Emulsion V.

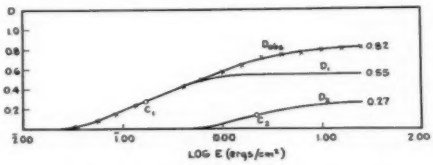


FIG. 16. XXX D_1+D_2. Emulsion VI.

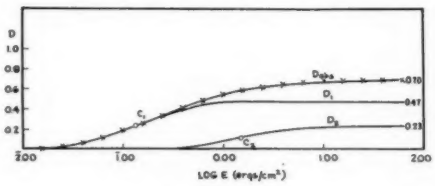


FIG. 17. Emulsion VII.

of the size-frequency distribution curve. Both, $\bar{a}$ and σ, will be quoted presently, but in the present connection we shall be concerned only with the average grain size $\bar{a}$.

The emulsions will, for brevity, be referred to as emulsions I, II to VII in ascending order of their

precipitation times. The corresponding values of $\bar{a}$ and σ (in μ^2), for the seven emulsions are:

Em.	I	II	III	IV	V	VI	VII
$\bar{a}=$	0.34	0.52	0.75	0.99	1.56	2.07	2.89
$\sigma=$	0.18	0.28	0.44	0.64	1.17	1.72	2.72

The characteristic curves of these seven emulsions, each for a 2-min. development in DK-50, are the upper solid curves, marked D_{obs}, in Figs. 11 to 17, corresponding to emulsions I to VII. The calculated sums D_1+D_2 are marked by crosses. The complete expressions found for D_1 and D_2 are inscribed in the figures; they contain the values of the proportionality coefficients c_1, c_2 explicitly. Thus, e.g., for the emulsion I,

$$D_1=1.54P(0.158E,\ 0.50),$$
$$D_2=0.76P(0.449E,\ 2.05).$$

C_1 and C_2 are the centers of symmetry of the component curves.

It will be seen from Figs. 11 to 17 that all the observed curves are represented throughout very closely by D_1+D_2.

From the position of the two centers C_1 and C_2, the sensitivities s_1 and s_2 are derived directly, as explained before. Their values are:

Em.	I	II	III	IV	V	VI	VII
$\bar{a}=$	0.34	0.52	0.75	0.99	1.56	2.07	2.89
$s_1=$	0.31	1.07	2.63	2.51	4.17	5.89	7.59
$s_2=$	0.22	0.18	0.41	0.32	0.50	0.46	0.68

While s_2 shows comparatively small and erratic variations, the sensitivity s_1 of the primary component of the emulsion is (apart from a slight upward jerk at $\bar{a}=0.75$) very nearly a *linear* function of the average grain size of the whole emulsion, steadily ascending with this grain size.

An equally close representation by the two-term formula has been obtained for the curves of the same seven emulsions found experimentally at a few other development times in DK-50 extending up to 8 min. and at six development times in ferrous oxalate ranging from 1 to 18 min. The behavior of the two sensitivities s_1, s_2, derived from the corresponding parameter values, has turned out to be very much the same as above. The peculiarities of these curves (and of their representation) being quite similar to those shown in Figs. 11 to 17, these curves need not be reproduced here.

The numerous examples given in this paper (with many others worked out, but omitted here for space economy) will suffice, perhaps, to show that the asymptotic formula $D=D_mP(y, r)$ and its superposition, $D=A_1P(y_1, r_1)+A_2P(y_2, r_2)$, even with the particular ratio $A_1:A_2=2:1$, may claim to be uniformly close, and easily handled, representatives of the characteristic curves of (respectively) non-sensitized and sensitized emulsions, so far, that is, pure silver bromide emulsions.

The somewhat intricate complications produced by the addition of iodide are now under investigation and their discussion must be relegated to a future opportunity.

My best thanks are due to Mr. A. P. H. Trivelli for the experimental material utilized in this paper and for some valuable suggestions.

APPENDIX

The following table of the function

$$P(y, r)=\tfrac{1}{2}\left\{1+\Phi\left(\frac{y-r}{(2y)^{\frac{1}{2}}}\right)\right\}$$

will be found convenient and sufficient for most or all practical applications:

TABLE I. $P(y, r)$

r \ y	r	$2r$	$4r$	$8r$	$16r$	$32r$	$64r$
0.25	0.500	.638	773	892	970	997	1.000
0.30	0.500	.651	794	912	980	999	1.000
0.35	0.500	.662	813	928	987	999	1.000
0.40	0.500	.673	829	941	991	1.000	1.000
0.45	0.500	.682	843	952	994	1.000	1.000
0.50	0.500	.692	856	960	996	1.000	1.000
0.55	0.500	.700	867	967	997	1.000	1.000
0.60	0.500	.708	877	973	998	1.000	1.000
0.65	0.500	.716	887	977	999	1.000	1.000
0.70	0.500	.723	895	981	999	1.000	1.000
0.75	0.500	.730	903	984	999	1.000	1.000
0.80	0.500	.736	910	987	1.000	1.000	1.000
0.85	0.500	.743	917	989	1.000	1.000	1.000
0.90	0.500	.749	923	991	1.000	1.000	1.000
0.95	0.500	.755	928	992	1.000	1.000	1.000
1.00	0.500	.760	933	993	1.000	1.000	1.000
1.10	0.500	.771	942	995	1.000		
1.20	0.500	.781	950	997	1.000		
1.30	0.500	.790	956	998	1.000		
1.40	0.500	.799	962	998	1.000		
1.50	0.500	.807	967	999	1.000		
1.60	0.500	.815	971	999	1.000		
1.70	0.500	.822	975	999	1.000		
1.80	0.500	.829	978	999_5	1.000		
1.90	0.500	.835	981	1.000	1.000		
2.00	0.500	.841	983	1.000	1.000		
2.20	0.500	.853	987	1.000			
2.40	0.500	.863	990	1.000			
2.60	0.500	.873	992	1.000			
2.80	0.500	.882	994	1.000			
3.00	0.500	.890	995	1.000			

A New Calculus for the Treatment of Optical Systems. IV.

R. CLARK JONES†
Research Laboratory, Polaroid Corporation, Cambridge, Massachusetts, and Research Laboratory of Physics, Harvard University, Cambridge, Massachusetts

(Received May 1, 1942)

Part IV is divided into two sections. The first is devoted to some additions to the general theory developed in Part I, and the second section to the derivation of the matrices representing two optical elements which were not treated in Parts II and III: (1) plates possessing circular dichroism, and (2) plates cut from crystals of such low symmetry that the principal axes of absorption and refraction are not parallel. In case (2), the discussion is limited to monoclinic and triclinic crystals which do not possess optical activity.

INTRODUCTION

IN the first three parts of this series which appeared together in a recent issue[1] of this Journal, the writer presented the outlines of a new calculus for the treatment of the type of optical system which contains plane-parallel plates of crystalline or other anisotropic material. The usefulness and power of the calculus was demonstrated by the applications in Parts II and III.

The purpose of the present paper is to make some extensions of the general theory which were suggested by members of the Research Laboratory of the Polaroid Corporation, and also to present further applications of the theory: we here derive the matrices representing (1) plates possessing circular dichroism and (2) plates cut from monoclinic and triclinic crystals.

EMENDATIONS TO THE GENERAL THEORY

Eigenvectors of a Matrix

In the case of a partial polarizer or a retardation plate, we know that plane polarized light with its plane parallel to either of the principal axes is transmitted through the plate without change in the state of polarization, although there may be a change in the intensity of the light. Similarly, right-handed and left-handed circularly polarized light suffer no change in their state of polarization in passing through a rotator or through a crystal exhibiting circular dichroism. In terms of the general theory developed in Part I, we see that in each of these cases just mentioned, the vector $\mathcal{E}$ representing the emergent light is equal to the vector* representing the entering light multiplied by some (in general complex) constant. A vector which has this property with respect to a given optical system or with respect to the matrix representing it, will be termed an *eigenvector* of the system or matrix.

The questions now arise, does every optical system have just two eigenvectors, and how may the eigenvectors of a given optical system be determined? We proceed to answer these questions.

As in I, let $\mathbf{M}^{(n)}$ be the matrix of the optical system. Then the condition that $\mathcal{E}_{(i)}$ be an eigenvector of $\mathbf{M}^{(n)}$ is

$$\mathbf{M}^{(n)}\mathcal{E}_{(i)}=d_i\mathcal{E}_{(i)}, \tag{1}$$

where d_i is a constant, at present undetermined, which we shall term the *eigenvalue* corresponding to the eigenvector $\mathcal{E}_{(i)}$.

The condition that (1) possess a non-vanishing solution for the two components of $\mathcal{E}_{(i)}$ is

$$\det\left[\mathbf{M}^{(n)}-d_i\right]=0 \tag{2}$$

or

$$(m_1-d_i)(m_2-d_i)=m_3m_4. \tag{3}$$

Let the two solutions of (3) be d_1 and d_2. Then from (1) we find that the corresponding eigenvectors are

$$\mathcal{E}_{(1)}\equiv\begin{pmatrix}a_1\\ b_1\end{pmatrix}\doteq\begin{pmatrix}m_1-d_2\\ m_3\end{pmatrix}\doteq\begin{pmatrix}m_4\\ m_2-d_2\end{pmatrix}\doteq\begin{pmatrix}d_1-m_2\\ m_3\end{pmatrix}\doteq\begin{pmatrix}m_4\\ d_1-m_1\end{pmatrix} \tag{4}$$

† Part of the work reported here was done while the writer held a National Scholarship at Harvard University; he is now a Member of the Technical Staff, Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

* The light vector $\mathcal{E}$ was mistakenly printed as a lower case epsilon in Parts I, II, and III. The error is here corrected.

and

$$\mathcal{E}_{(2)} \equiv \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \doteq \begin{pmatrix} m_1 - d_1 \\ m_3 \end{pmatrix} \doteq \begin{pmatrix} m_4 \\ m_2 - d_1 \end{pmatrix} \doteq \begin{pmatrix} d_2 - m_2 \\ m_3 \end{pmatrix} \doteq \begin{pmatrix} m_4 \\ d_2 - m_1 \end{pmatrix}, \quad (5)$$

where the symbol $\doteq$ should be read "equal except for a constant factor." It is, of course, not possible to define the eigenvector uniquely, since if $\mathcal{E}_{(i)}$ is an eigenvector, then so is any multiple of $\mathcal{E}_{(i)}$.

Up to this point, we have found that every optical system has, at most, two eigenvectors, and we have found general expressions for these eigenvectors. We must now determine whether $\mathcal{E}_{(1)}$ and $\mathcal{E}_{(2)}$ are linearly independent, since if they are merely multiples of one another, they are essentially the same eigenvector.

We shall now show that the necessary and sufficient condition for $\mathbf{M}^{(n)}$ to possess two and only two** linearly independent eigenvectors, is that $\mathbf{M}^{(n)}$ be diagonalizable—that is to say, that a matrix $\mathbf{T}$ exists such that

$$\mathbf{T}^{-1}\mathbf{M}^{(n)}\mathbf{T} = \mathbf{D} \equiv \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}. \quad (6)$$

We first point out the obvious fact that the linear independence of $\mathcal{E}_{(1)}$ and $\mathcal{E}_{(2)}$ is equivalent to the condition that

$$\Delta \equiv a_1 b_2 - a_2 b_1 \quad (7)$$

be different from zero. We now define the matrices

$$\mathbf{T} \equiv \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}; \quad \mathbf{T}^{-1} \equiv \Delta^{-1} \begin{pmatrix} b_1 & -a_2 \\ -b_1 & a_1 \end{pmatrix} \quad (8)$$

and by actual substitution of (4), (5), and (8) in (6), we find that (6) is satisfied. We have thus found a matrix which diagonalizes $\mathbf{M}^{(n)}$. Although the condition (6) does not determine $\mathbf{T}$ uniquely, detailed examination confirms what is suggested by (8): when Δ is zero, no matrix $\mathbf{T}$ exists which satisfies (8). We have thus proved the assertion at the beginning of this paragraph.

** In making this statement we are ignoring the degenerate case in which every vector is an eigenvector of the matrix, as is the case with the matrix representing an isotropic crystal.

The Matrix with Given Eigenvectors and Given Eigenvalues

Suppose that a given optical system is found by experiment to have the eigenvectors

$$\mathcal{E}_{(1)} \equiv \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}; \quad \mathcal{E}_{(2)} \equiv \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \quad (9)$$

with the corresponding eigenvalues d_1 and d_2. The matrix of the optical system is then uniquely determined, and is given by

$$\mathbf{M}^{(n)} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1} \equiv \Delta^{-1} \begin{pmatrix} d_1 a_1 b_2 - d_2 a_2 b_1 & -(d_1 - d_2) a_1 a_2 \\ (d_1 - d_2) b_1 b_2 & d_2 a_1 b_2 - d_1 a_2 b_1 \end{pmatrix}. \quad (10)$$

This statement may be confirmed by subjecting the matrix (10) to the analysis of the last section.

Other Representations of the Light Vector

In the development of the general theory in Part I, the state of polarization of the light was expressed by stating the amplitudes and phases of the two Cartesian components of the plane wave. That is to say, the light vector was expressed as a linear combination of the two *basis vectors*

$$\mathcal{E}_{(1)} \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \quad \mathcal{E}_{(2)} \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (11)$$

in the manner

$$\mathcal{E} \equiv \begin{pmatrix} E_x \\ E_y \end{pmatrix} \equiv E_x \mathcal{E}_{(1)} + E_y \mathcal{E}_{(2)}. \quad (12)$$

The writer is indebted to Dr. Martin Grabau for pointing out that it is sometimes expedient to use basis vectors which are not Cartesian—for example, in dealing with optical systems in which most of the elements are rotators or crystals exhibiting circular dichroism, it is clearly more convenient to express the light vector as the sum of its two circularly polarized components.

We shall now develop some of the relations which hold when the basis vectors are chosen arbitrarily. Let $\mathcal{E}_{(1)}$ and $\mathcal{E}_{(2)}$ be two arbitrary and linearly independent vectors, with Cartesian components as defined in (9). For the time being, the vectors $\mathcal{E}_{(1)}$ and $\mathcal{E}_{(2)}$ are *not* to be thought of as eigenvectors. Let us define a new light vector

in the manner

$$\mathfrak{F} \equiv \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} \equiv F_1 \mathcal{E}_{(1)} + F_2 \mathcal{E}_{(2)}. \tag{13}$$

Evidently, the vector $\mathfrak{F}$ is defined only with respect to some given pair of basis vectors.

From (11), (12), and (13), we find that the relations between $\mathcal{E}$ and $\mathfrak{F}$ are

$$\mathcal{E} = \mathbf{T}\mathfrak{F}; \quad \mathfrak{F} = \mathbf{T}^{-1}\mathcal{E}, \tag{14}$$

where $\mathbf{T}$ is defined by (8). If $\mathbf{M}^{(n)}$ is the matrix of an optical system in the $\mathcal{E}$-representation:

$$\mathcal{E}_n = \mathbf{M}^{(n)}\mathcal{E}_0, \tag{15}$$

then $\mathbf{T}^{-1}\mathbf{M}^{(n)}\mathbf{T}$ is the matrix in the $\mathfrak{F}$-representation:

$$\mathfrak{F}_n = \mathbf{T}^{-1}\mathbf{M}^{(n)}\mathbf{T}\mathfrak{F}_0. \tag{16}$$

Conversely, if $\mathbf{M}^{(n)}$ is the matrix in the $\mathfrak{F}$-representation, then $\mathbf{T}\mathbf{M}^{(n)}\mathbf{T}^{-1}$ is the matrix in the $\mathcal{E}$-representation.

If now we use as basis vectors the eigenvectors of $\mathbf{M}^{(n)}$, where $\mathbf{M}^{(n)}$ is the matrix of an optical system in the $\mathcal{E}$-representation, then the matrix of the optical system in the $\mathfrak{F}$-representation is the diagonal matrix $\mathbf{D}$, as we may prove by comparing (16) with (6). It is this fact which sometimes makes it convenient to use as basis vectors the eigenvectors of the optical system or the eigenvectors of some element of the system.

At this point, it is desirable to introduce a few definitions. A vector $\mathcal{E}_{(1)}$ is *normal* if

$$\bar{\mathcal{E}}_{(1)}\mathcal{E}_{(1)}^* \equiv aa^* + bb^* = 1 \tag{17}$$

and two vectors are *orthogonal* with respect to one another if

$$\bar{\mathcal{E}}_{(1)}\mathcal{E}_{(2)}^* \equiv a_1a_2^* + b_1b_2^* = 0. \tag{18}$$

Two vectors are *orthonormal* if each of them is normal, and if they are orthogonal. The condition that the eigenvectors of a matrix be normal involves no loss in generality, but the requirement that they be orthogonal places a restriction on the nature of the matrix. A sufficient condition that the eigenvectors of $\mathbf{M}$ be orthogonal is that $\mathbf{M}$ be either Hermitian or unitary, and a necessary condition is that the absolute values of m_3 and m_4 be equal. The necessary *and* sufficient condition that the eigenvectors of $\mathbf{M}$ be orthogonal is that $\mathbf{M}$ commute with its Hermitian adjoint:[3]

$$\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger\mathbf{M}.$$

It was pointed out in Part I that the intensity, of the light is proportional to $\bar{\mathcal{E}}\mathcal{E}^*$. If and only if the basis vectors of the $\mathfrak{F}$-representation are orthonormal, however, do we have

$$\bar{\mathcal{E}}\mathcal{E}^* = \bar{\mathfrak{F}}\mathfrak{F}^*. \tag{19}$$

Equation (19) is a special case of Plancherel's theorem.[2] More generally, if $\mathcal{E}_j$ and $\mathcal{E}_k$ are any two vectors and if the basis vectors are orthonormal, we have

$$\bar{\mathcal{E}}_j\mathcal{E}_k^* = \bar{\mathfrak{F}}_j\mathfrak{F}_k^*. \tag{20}$$

If the basis vectors are orthonormal, it is no real loss in generality to take them to be

$$\mathcal{E}_{(1)} \equiv \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}; \quad \mathcal{E}_{(2)} \equiv \begin{pmatrix} -b_1^* \\ a_1^* \end{pmatrix} \tag{21}$$

with

$$a_1a_1^* + b_1b_1^* = 1, \tag{22}$$

in which case Δ is equal to unity.

Partially Polarized Light

Up to this point, it has always been assumed that the light is in a definite state of polarization. Nearly all of the ordinary sources of light, however, yield light which is almost completely unpolarized. In the following treatment of partially polarized light, it will be assumed that light is passing through the optical system in only one direction, so that it is proper to infer that the intensity, or energy flux vector, is parallel to the axis of the system, and that its length is proportional to the square of the length of the light vector.

Unpolarized light of unit intensity may be described as the combination of two linearly polarized components, one parallel to the x axis, and the other parallel to the y axis, each of intensity one-half, and with the added stipulation that all values of the relative phase of the two components are equally likely, so that, when one computes the total intensity, it is necessary to average over the relative phase. There is, however, nothing sacred about light polarized along the axes of the coordinate system, nor is there anything sacred about linearly polarized light. More generally, unpolarized light of unit in-

tensity may be regarded as the superposition of any two orthogonal components, each of intensity one-half, and with all values of the relative phase equally likely.

In the following discussion we make the assumption, convenient for ease of presentation, that a normal light vector has an intensity of one-half. Such a correspondence may always be obtained by suitable choice of the units in which energy and the electric vector are expressed. Let $\mathcal{E}_{01}$ and $\mathcal{E}_{02}$ be the two orthonormal components in terms of which unpolarized light of unit intensity is described. Then, after passing through an optical system with the matrix $\mathbf{M}^{(n)}$, the corresponding components of the emergent light will be

$$\mathcal{E}_{n1}=\mathbf{M}^{(n)}\mathcal{E}_{01}$$

and (23)

$$\mathcal{E}_{n2}=\mathbf{M}^{(n)}\mathcal{E}_{02}.$$

In general, $\mathcal{E}_{n1}$ and $\mathcal{E}_{n2}$ will be neither normal nor orthogonal. The emergent light may be described as the combination of the vectors $\mathcal{E}_{n1}$ and $\mathcal{E}_{n2}$, with the understanding that all values of the relative phase are equally likely. If one wishes to know the intensity of the transmitted light, one simply finds the intensity of $\mathcal{E}_{n1}$ and of $\mathcal{E}_{n2}$, and adds these two intensities. Or, more generally, if one wishes to know the intensity of a specified component in the emergent light, one finds the intensity of the given component in $\mathcal{E}_{n1}$ and $\mathcal{E}_{n2}$, and adds.

The result of computing the intensity of the transmitted light, or the intensity of any component of it, is, as it must be, independent of the arbitrary choice of the orthonormal vectors $\mathcal{E}_{01}$ and $\mathcal{E}_{02}$. The intensity transmission factor for initially unpolarized light of an optical system with the matrix $\mathbf{M}^{(n)}$ is easily shown to be

$$(|\mathbf{M}^{(n)}\mathcal{E}_{01}|^2+|\mathbf{M}^{(n)}\mathcal{E}_{02}|^2)=\tfrac{1}{2}(|m_1|^2+|m_2|^2+|m_3|^2+|m_4|^2). \quad (24)$$

If $\mathcal{E}_{n1}$ and $\mathcal{E}_{n2}$ have the same absolute value and if they are orthogonal, the emergent light is also unpolarized. If they differ in intensity, or if they are not orthogonal, the emergent light is said to be *partially polarized*. More specially, the emergent light is *completely polarized* if either $\mathcal{E}_{n1}$ or $\mathcal{E}_{n2}$ is zero, or if they are linearly dependent.

Further Discussion of Γ

In Part II (p. 496), we introduced a quantity $\Gamma(\mathbf{M})$, which was defined as the maximum value of $|\mathbf{M}\mathcal{E}|/|\mathcal{E}|$ with respect to variations of the two components of $\mathcal{E}$. We see that Γ^2 is equal to the maximum value of the intensity transmission factor of the optical system represented by $\mathbf{M}$. If we denote the intensity transmission factor by T, we have

$$T\equiv|\mathbf{M}\mathcal{E}|^2/|\mathcal{E}|^2. \quad (25)$$

In the following paragraphs, we shall find explicit expressions for the maximum value of T, and also for the minimum value, for any matrix $\mathbf{M}$.

For the present purpose, it is no loss of generality to assume that $\mathcal{E}$ has the following normal form:

$$\mathcal{E}=\begin{pmatrix} e^{i\gamma}\sin\theta \\ e^{-i\gamma}\cos\theta \end{pmatrix}. \quad (26)$$

We now find that T is given explicitly by the following expression:

$$\begin{aligned} T\equiv|\mathbf{M}\mathcal{E}|^2=&(|m_1|^2+|m_3|^2)\sin^2\theta \\ &+(|m_2|^2+|m_4|^2)\cos^2\theta \\ &+[(m_1m_4^*+m_2^*m_3)e^{2i\gamma} \\ &+(m_1^*m_4+m_2m_3^*)e^{-2i\gamma}]\sin\theta\cos\theta. \end{aligned} \quad (27)$$

In order to find the maximum value of T, and thus the value of Γ, we must assign to γ and θ the values which make T stationary with respect to variations in γ and θ. These values are found to be

$$\gamma=\frac{1}{2}\arg(m_1^*m_4+m_2m_3^*), \quad (28)$$

$$\tan 2\theta=\frac{2|m_1^*m_4+m_2m_3^*|}{|m_2|^2+|m_4|^2-|m_1|^2-|m_3|^2}, \quad (29)$$

whence

$$\begin{aligned} T_{\text{stat}}=&\tfrac{1}{2}(|m_1|^2+|m_2|^2+|m_3|^2+|m_4|^2 \\ &+\tfrac{1}{2}(|m_2|^2+|m_4|^2-|m_1|^2-|m_2|^2)\sec 2\theta, \end{aligned} \quad (30)$$

where θ is understood to have the value (29), or

$$\begin{aligned} T_{\text{stat}}=&\tfrac{1}{2}(|m_1|^2+|m_2|^2+|m_3|^2+|m_4|^2) \\ &\pm\tfrac{1}{2}[(|m_2|^2+|m_4|^2-|m_1|^2-|m_3|^2)^2 \\ &+4|m_1m_4^*+m_2^*m_3|^2]^{\frac{1}{2}} \end{aligned} \quad (31)$$

if θ is eliminated.

It is clear the $T_{\max}$ is to be associated with the plus sign in (31). We then have $\Gamma=(T_{\max})^{\frac{1}{2}}$.

By (29), we see that the two values of θ which yield the maximum and minimum values of T differ by 90°, and from this fact it follows at once that the vector $\mathcal{E}_{max}$ which maximizes T is orthogonal to the vector $\mathcal{E}_{min}$ which minimizes T. It should be noted, however, that in general $\mathcal{E}_{max}$ and $\mathcal{E}_{min}$ are *not* eigenvectors of the matrix $\mathbf{M}$; if and only if the matrix possesses orthogonal eigenvectors will $\mathcal{E}_{max}$ and $\mathcal{E}_{min}$ be the eigenvectors of the matrix.

It is interesting to note, by comparison of (31) with (24), that the arithmetic mean of T_{max} and T_{min} is equal to the transmission factor for unpolarized light.

It has been pointed out earlier in this part (cf. paragraph following Eq. (16)) that if one uses as basis vectors the eigenvectors of $\mathbf{M}$, then in the $\mathfrak{F}$-representation, the matrix $\mathbf{M}$ becomes diagonal, with the diagonal elements d_1 and d_2. The reader should be warned against the error of concluding from this fact, that $\Gamma(\mathbf{M})$ is equal to the larger of $|d_1|$ and $|d_2|$. It follows from the last two paragraphs that $\Gamma(\mathbf{M})$ is equal to the larger of $|d_1|$ and $|d_2|$ if and only if the eigenvectors of $\mathbf{M}$ are orthogonal.

APPLICATIONS

Circular Birefringence and Circular Dichroism

Fresnel's phenomenological explanation of the optical activity of quartz was that the indices of refraction of quartz were different for right and left circularly polarized light. It is evident physically that such a difference will produce a rotation of the plane of linearly polarized light.

In Part I we found that the matrix of a rotator was $\mathbf{S}(\tilde{\omega})$ by making use of the fact that a rotator produced a rotation of the plane of linearly polarized light which was independent of the angle of the plane of polarization. It is interesting to derive this matrix also by the use of Fresnel's assumption. By assumption, the normal eigenvectors of a rotator are

$$\mathcal{E}_{(1)}=2^{-\frac{1}{2}}\begin{pmatrix}1\\ i\end{pmatrix},\quad \mathcal{E}_{(2)}=2^{-\frac{1}{2}}\begin{pmatrix}-1\\ i\end{pmatrix} \tag{32}$$

and we shall indicate the relative phase retardation of $2\tilde{\omega}$ by writing

$$d_1=e^{i\tilde{\omega}},\quad d_2=e^{-i\tilde{\omega}}. \tag{33}$$

By substituting these expressions in (10), we find at once that the matrix of a rotator is

$$\begin{pmatrix}\cos\tilde{\omega} & -\sin\tilde{\omega}\\ \sin\tilde{\omega} & \cos\tilde{\omega}\end{pmatrix}\equiv\mathbf{S}(\tilde{\omega}). \tag{34}$$

The fact that we obtain the same matrix by considering a rotator *qua* a rotator, and as a crystal possessing circular birefringence, provides a mathematical proof of the physical equivalence of the two points of view.

Circular dichroism differs from circular birefringence only in that the two components are differentially absorbed, rather than differentially retarded. In the case of pure circular dichroism, therefore, we take the eigenvalues to be real. From (10) and (32), we then find for the matrix

$$\begin{pmatrix}\frac{1}{2}(d_1+d_2) & \frac{1}{2}i(d_1-d_2)\\ -\frac{1}{2}i(d_1-d_2) & \frac{1}{2}(d_1+d_2)\end{pmatrix}. \tag{35}$$

Circular dichroism is of interest primarily in connection with liquids, although a measurement of its magnitude in crystalline nickel sulphate has been reported recently.[4]

Elliptical Birefringence and Elliptical Dichroism

The treatment of elliptical birefringence and dichroism is exactly analogous to the foregoing treatment. If the eigenvectors are orthogonal, the matrix for the case of elliptical birefringence turns out to be the general unitary matrix, and we have seen that this matrix is the most general matrix which one encounters with optical elements which have no absorption. A matrix with the eigenvalues (33) and with non-orthogonal eigenvectors has a value of Γ which is greater than unity; such a matrix is therefore not physically realizable.

Crystals of Low Symmetry: Monoclinic and Triclinic

Monoclinic and triclinic crystals are crystals of such low symmetry that the principal axes of absorption are not parallel to the principal axes of refraction. In this section we shall derive by a new approach the matrix which corresponds to a crystalline plate cut from such crystals. Our treatment will be restricted to those crystals which do not possess optical activity—that is to

say, to those of the crystal classes C_s and C_{2h} (monoclinic), and C_i (triclinic).

Our explicit treatment will be restricted to the case in which the crystalline plate is cut from a monoclinic crystal with the symmetry axis normal to the surface of the plate. At the conclusion of the derivation, however, we shall show that the results are also valid for triclinic crystals.

The general equations for the propagation of plane waves of light through triclinic crystals have been developed by the use of standard electromagnetic theory.[5] The derivation by this means is entirely rigorous, but it is tedious, and the resulting equations, in the general case, are complicated in form and yield little or no physical insight.

The present derivation of the matrix representing a monoclinic crystal without optical activity and with its axis normal to the surface of the plate is based on the observation that such a crystal is physically equivalent to an optical system whose odd numbered elements are identical partial polarizers each with the same orientation ω_P and whose even numbered elements are identical retardation plates with the orientation ω_G, in the limit in which the number of plates becomes infinite and their thickness zero with the total thickness of the optical system remaining constant. Indeed, the writer finds that this method of representing a monoclinic crystal is the only one which provides him any physical insight into its optical behavior.

In the following derivation, we shall first determine the matrix corresponding to a finite number of pairs of partial polarizers and retardation plates, and we shall then pass to the indicated limit.

The matrix of a retardation plate with the orientation ω_G is

$$\mathbf{S}(\omega_G)\mathbf{G}\mathbf{S}(-\omega_G), \tag{36}$$

where

$$\mathbf{G}\equiv\begin{pmatrix} e^{i\gamma} & 0 \\ 0 & e^{-i\gamma} \end{pmatrix}. \tag{37}$$

Similarly, the matrix of a partial polarizer with the orientation ω_P is

$$\mathbf{S}(\omega_P)\mathbf{P}\mathbf{S}(-\omega_P), \tag{38}$$

where

$$\mathbf{P}\equiv\begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}=\begin{pmatrix} \kappa e^{\xi} & 0 \\ 0 & \kappa e^{-\xi} \end{pmatrix} \tag{39}$$

$$\left.\begin{aligned} &0\leq p_1\leq 1,\\ &0\leq p_2\leq 1,\\ &\kappa=(p_1p_2)^{\frac{1}{2}},\\ &\xi=\tfrac{1}{2}\log(p_1/p_2). \end{aligned}\right\} \tag{40}$$

If we suppose that the light passes first through the partial polarizer, the matrix of the combination of the partial polarizer and retardation plate is then

$$\mathbf{B}\equiv\mathbf{S}(\omega_G)\mathbf{G}\mathbf{S}(-\omega_G)\mathbf{S}(\omega_P)\mathbf{P}\mathbf{S}(-\omega_P) \tag{41}$$

and the matrix representing the optical system composed of n such pairs, one after the other, is $\mathbf{B}^n$, according to Eq. (15, I).

It will materially reduce the algebraic complexity of the resulting equations, however, if instead of determining $\mathbf{B}^n$ directly, we first determine the nth power of

$$\mathbf{A}\equiv\mathbf{S}(-\omega_G)\mathbf{B}\mathbf{S}(\omega_G)=\mathbf{G}\mathbf{S}(\omega)\mathbf{P}\mathbf{S}(-\omega), \tag{42}$$

where

$$\omega\equiv\omega_P-\omega_G \tag{43}$$

and then determine $\mathbf{B}^n$ from $\mathbf{A}^n$ by the simple relation

$$\mathbf{B}^n=\mathbf{S}(\omega_G)\mathbf{A}^n\mathbf{S}(-\omega_G). \tag{44}$$

The physical significance of this mathematical simplification is that we are first rotating our coordinate axes so that they are parallel to the principal axes of the retardation plate; the transformation (44) then corresponds to rotating the coordinate axes back to their original position.

From Eq. (42) we find that the matrix elements of $\mathbf{A}$ are

$$\left.\begin{aligned} a_1&=\kappa e^{i\gamma}(\cosh\xi+\cos 2\omega\sinh\xi),\\ a_2&=\kappa e^{-i\gamma}(\cosh\xi-\cos 2\omega\sinh\xi),\\ a_3&=\kappa e^{-i\gamma}\sin 2\omega\sinh\xi,\\ a_4&=\kappa e^{i\gamma}\sin 2\omega\sinh\xi. \end{aligned}\right\} \tag{45}$$

In order to obtain the nth power of the matrix $\mathbf{A}$ by the use of Eqs. (8, III), it is first necessary to determine the characteristic values of $\mathbf{A}$. By (3),

we find

$$\left.\begin{aligned} d_1, d_2 &= \alpha \pm (\alpha^2-1)^{\frac{1}{2}} \\ &= (1+\beta^2)^{\frac{1}{2}} \pm \beta \\ &= \alpha \pm \beta, \end{aligned}\right\} \tag{46}$$

where

$$\left.\begin{aligned} \alpha &\equiv \cosh \xi \cos \gamma + i \cos 2\omega \sinh \xi \sin \gamma, \\ \beta &\equiv (\alpha^2-1)^{\frac{1}{2}}, \end{aligned}\right\} \tag{47}$$

and where we shall in the following correlate d_1 with the plus sign in (46).

As a convenient abbreviation, we take $\mathbf{A}' \equiv \mathbf{A}^n$. We then have

$$d_1' = (\alpha+\beta)^n, \quad d_2' = (\alpha-\beta)^n \tag{48}$$

and substitution of (45) and (48) in (8, III) yields the explicit expressions for the matrix elements of $\mathbf{A}' \equiv \mathbf{A}^n$.

As the second step in our determination of the matrix representing a monoclinic crystal, we must take the limit of $\mathbf{A}'$ as n becomes infinite under the condition that γ_0, ξ_0, and κ_0 are held constant:

$$\gamma_0 = n\gamma, \quad \xi_0 = n\xi, \quad \kappa_0 = \kappa^n. \tag{49}$$

We find

$$\left.\begin{aligned} \lim a_1' &= \lim\, (\kappa^n/2\beta)[(a_1-1+\beta)(1+\beta)^n \\ &\qquad -(a_1-1-\beta)(1-\beta)^n \\ &= \kappa_0(\cosh z + \Psi_1 \sinh z), \\ \lim a_2' &= \kappa_0(\cosh z - \Psi_1 \sinh z), \\ \lim a_3' &= \lim a_4' = \kappa_0 \Psi_2 \sinh z, \end{aligned}\right\} \tag{50}$$

where

$$\left.\begin{aligned} z^2 &\equiv \xi_0^2 - \gamma_0^2 + 2ix_0\gamma_0 \cos 2\omega = \lim n^2\beta, \\ \Psi_1 &\equiv (\xi_0 \cos 2\omega + i\gamma_0)/z, \\ \Psi_2 &\equiv (\xi_0 \sin 2\omega)/z. \end{aligned}\right\} \tag{51}$$

It is easily confirmed that the matrix elements of $\lim \mathbf{A}'$ are independent of which of the two possible values of z is used, as long as the same root is used throughout.

Finally, we obtain the explicit expression for $\mathbf{\Theta}$, the matrix representing a monoclinic crystal whose two pairs of principal axes have the orientations ω_G and ω_P:

$$\begin{aligned} \lim \mathbf{B}^n &= \mathbf{S}(\omega_G)(\lim \mathbf{A}')\mathbf{S}(-\omega_G) \\ &= \mathbf{\Theta} \equiv \begin{pmatrix} \kappa_0(\cosh z + \Gamma_1 \sinh z) & \kappa_0\Gamma_2 \sinh z \\ \kappa_0\Gamma_2 \sinh z & \kappa_0(\cosh z - \Gamma_1 \sinh z) \end{pmatrix} \end{aligned} \tag{52}$$

in which we have used the abbreviations

$$\left.\begin{aligned} z^2 &\equiv \xi_0^2 - \gamma_0^2 + 2i\xi_0\gamma_0 \cos 2(\omega_P - \omega_G), \\ \Gamma_1 &\equiv (\xi_0 \cos 2\omega_P + i\gamma_0 \cos 2\omega_G)/z, \\ \Gamma_2 &\equiv (\xi_0 \sin 2\omega_P + i\gamma_0 \sin 2\omega_G)/z. \end{aligned}\right\} \tag{53}$$

In this case also, it matters not which of the two possible values of z is used.

If the optical system represented by $\mathbf{\Theta}$ has a thickness t, it is clear by comparison with Eq. (2, I) that the difference of the two principal indices of refraction is given by

$$n_{y'} - n_{x'} = \lambda\gamma_0/\pi t \tag{54}$$

and that the two principal extinction coefficients are

$$\left.\begin{aligned} k_{x'} &= \lambda(\log \kappa_0 + \xi_0)/2\pi t, \\ k_{y'} &= \lambda(\log \kappa_0 - \xi_0)/2\pi t. \end{aligned}\right\} \tag{55}$$

These relations serve to evaluate ξ_0, γ_0, and κ_0 in terms of the principal optical coefficients of the crystal.

The two eigenvectors of $\mathbf{\Theta}$ are

$$\mathcal{E}_{(1)} \doteq \begin{pmatrix} 1+\Gamma_1 \\ \Gamma_2 \end{pmatrix} \doteq \begin{pmatrix} \Gamma_2 \\ 1-\Gamma_1 \end{pmatrix} \tag{56}$$

and

$$\mathcal{E}_{(2)} \doteq \begin{pmatrix} \Gamma_1 - 1 \\ \Gamma_2 \end{pmatrix} \doteq \begin{pmatrix} -\Gamma_2 \\ 1+\Gamma_1 \end{pmatrix}. \tag{57}$$

Because of the identity,

$$\Gamma_1^2 + \Gamma_2^2 \equiv 1 \tag{58}$$

the eigenvectors satisfy the relation $\bar{\mathcal{E}}_{(1)}\mathcal{E}_{(2)} = 0$, but do not satisfy the orthogonality condition $\bar{\mathcal{E}}_{(1)}\mathcal{E}_{(2)}{}^* = 0$. The eigenvectors of $\mathbf{\Theta}$ are therefore orthogonal if and only if $\mathcal{E}_{(2)}$ and $\mathcal{E}_{(2)}{}^*$ are linearly dependent, and this is not true in general.

Triclinic Crystals

Up to this point, our detailed considerations have related exclusively to monoclinic crystals.

There is a sense, however, in which the results are also valid for triclinic crystals, as we shall now show.

It is characteristic of all crystals without optical activity that their effect on light passing through them is the same whether the light is being propagated in one direction or in the reverse direction. From a comparison of Eqs. (14) and (23) of Part I, it follows at once that the matrix representing a plate cut from a crystal without optical activity must be symmetrical.

Secondly, it is not difficult to show that *any* symmetrical matrix with a determinant which is real and non-negative may be written in the form Θ with the quantities κ_0, γ_0, ξ_0, ω_G, and ω_P all real. The proof is similar to those used in Part II. It therefore follows from this paragraph and the last that the matrix representing any triclinic crystal without optical activity may also be written in the form Θ.

In the triclinic case, however, the quantities κ_0, γ_0, ξ_0, ω_G, and ω_P are not so directly related to the principal coefficients as in the monoclinic case (cf. Eqs. (54) and (55)). It is convenient to consider these quantities as being those of the monoclinic crystal which is equivalent to the triclinic crystal under discussion. The equivalence is valid only for a plate cut from the triclinic crystal at a given angle.

ACKNOWLEDGMENT

The writer would like to acknowledge the helpful discussion of various members of the Polaroid Corporation. He is particularly indebted to Dr. Cutler D. West and to Dr. Martin Grabau for their helpful suggestions.

REFERENCES

(1) R. Clark Jones, J. Opt. Soc. Am. **31**, 488 (1941); Henry Hurwitz, Jr., and R. Clark Jones, J. Opt. Soc. Am. **31**, 493 (1941); R. Clark Jones, J. Opt. Soc. Am. **31**, 500 (1941).
(2) M. Plancherel, Rend. di Palermo **30**, 289 (1910).
(3) For the definition of $\mathbf{M}^{\dagger}$, see the footnote on p. 492 in Part I.
(4) L. R. Ingersoll, P. Rudnick, F. G. Slack, and M. Underwood, Phys. Rev. **57**, 1145 (1940).
(5) M. Berek, Fortschr., Mineral. Krist. Petrog. **22**, 1–104 (1937).

Book Review

Motivation and Visual Factors. Individual Studies of College Students. I. E. BENDER, H. A. IMUS, J. W. M. ROTHNEY, C. KEMPLE AND M. R. ENGLAND. Pp. xix+369. Dartmouth College Publications, Hanover, New Hampshire, 1942. Price $4.00.

This work is based on individual case studies of a considerable number of college men which were collected at the Dartmouth Eye Institute. In the beginning, the hope was to secure data from the entire class of 1940 at Dartmouth College, but it became necessary to narrow the scope of the program to 124 cases. The present volume reports on these data.

The purpose and general plan of the study are discussed, and the methods to be utilized are reviewed in four opening chapters (pp. 1–57). The major portion of the work (Chapter V, pp. 59–316) contains detailed psycho-portraits and summaries of twenty subjects selected to typify five different kinds of visual classification. Summary and general conclusions, supplementary data, glossary, and subject and name indices complete the study (pp. 317–369).

The main contents of the volume are descriptive materials concerning the goals which college men hold in view, summaries of measurements on visual capacities in the same individuals, and comparisons and interpretations based on these data. Information concerning motivation was obtained through various conventional procedures—interviews, projective techniques, autobiographies, personality questionnaires, behavior observations, ratings by associates, medical health records, reading tests, psychological examinations, tests of factual knowledge, etc. Psychophysical measurements were made as regards acuity, corneal astigmatism, refractive errors, heterophoria, peripheral fusion, aniseikonia, ocular dominance, color aptitude, and a rating was assigned as regards visual efficiency.

The central aim of the study was to ascertain whether, and possibly to what extent, differences in visual capacity influence the dynamic motivational pattern of the individual. The qualitative evidence submitted indicates that such visual handicaps as are found in men admitted to college do affect motivation, although in no regular way. In some cases, they may provide the excuse for failure; in others, the incentive to compensation; and, in still others, they seem to have no effect at all. "The final conclusion of this study is that . . . the particular motivational pattern of the individual is the principal factor that determines the nature of the individual's adjustment to his visual defects or their correction" (pp. 323–324).

Quantitative relationships were attempted in terms of the method of correlation. Although a correlation coefficient that is statistically significant was found to obtain between specific visual factors,—e.g., refractive error and distant aniseikonia, distant aniseikonia and near aniseikonia—the outcome of the attempt to establish quantitative relations among various aspects of motivation and different visual factors was not so successful. Ratings regarding visual efficiency and scores attained in scholastic and/or performance tests did not yield relationships that were statistically significant. Visual defects also did not correlate significantly with achievement in college or scores on psychological tests. The general conclusion is that no significant relationship exists between visual efficiency and academic grades or between visual efficiency and tested abilities.

Although this general conclusion may ultimately be acceptably substantiated, the reader is to be cautioned, at least for the present, against accepting a hasty judgment in regard to the non-existence of a positive relationship between aspects of motivation and visual capacity. Reasons for this reservation are several. (1) Quantitative relationships were attempted in this study in respect of only a limited number of aspects of motivation and visual factors. (2) It is acknowledged by the authors that differentiation as regards visual factors is not so great among college men as in an unselected population. (3) The rating scale employed in this investigation probably did not allow sufficient differentiation as regards visual efficiency. Although it provided a range from 0–5, the persons rated all fell within the limits of 2–5, and the 20 cases specially treated in Chapter V fell within the narrow bounds of 3.5–5. One of the three raters assigned no efficiency value lower than 3.5. (4) From the standpoint of method, greater effort might be made to nullify or to minimize weaknesses known to exist in the methods used. The methods utilized in the study of motivation are perhaps most in need of improvement, however, the reviewer is also inclined to question the adequacy of certain of the visual determinations that were made, e.g., that regarding color aptitude.

By way of general appraisal, this work renders an important service in bringing into focus the problem of determining relations among complex variables such as those constituting human personality. Although so far the problem of differentiating quantitative relations in human personality has proved rather refractory, this is owing in part, to the intrinsic complexity of the problem, and in part, to the need for improving the methods employed. The present work will serve to stimulate other workers to the task of fashioning a more adequate scientific approach to the problem of isolating the various functional components of personality and establishing quantitative relations among them.

MICHAEL J. ZIGLER
Wellesley College